
**TRIBHUVAN UNIVERSITY**

**INSTITUTE OF ENGINEERING**

**PULCHOWK CAMPUS**

**THESIS NO.: T03/076**

# Modelling Pedestrian-Vehicle Conflict and Severity at Uncontrolled Midblock Crossings Inside Kathmandu Valley

**by**

**Ashish Banstola**

**A THESIS**

**SUBMITTED TO THE DEPARTMENT OF CIVIL ENGINEERING**

**IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE**

**DEGREE OF MASTER OF SCIENCE IN TRANSPORTATION ENGINEERING**

**DEPARTMENT OF CIVIL ENGINEERING**

**LALITPUR, NEPAL**

**APRIL, 2025**

# COPYRIGHT

**TRIBHUVAN UNIVERSITY**

**INSTITUTE OF ENGINEERING**

**PULCHOWK CAMPUS**

**DEPARTMENT OF CIVIL ENGINEERING**

The undersigned certify that they have read and recommended to Institute of Engineering for acceptance, a thesis entitled **"Modelling Pedestrian-Vehicle Conflict and Severity at Uncontrolled Midblock Crossings Inside Kathmandu Valley"** submitted by Ashish Banstola in partial fulfillment of the requirement for degree of Master of Science in Transportation Engineering.

.....

Supervisor,
Asst. Prof.Dr. Pradeep Kumar Shrestha
Department of Civil Engineering
Institute of Engineering

.....

External Examiner,
Saroj Kumar Pradhan

.....

[redacted] Asst.Prof. Anil Marsani
Coordinator, M.Sc. in Transportation Engineering
Department of Civil Engineering

Date: 16 April 2025

# ABSTRACT

Pedestrian safety at uncontrolled urban midblock crossings is a critical prerequisite for sustainable urban transport. Evaluation of factors affecting pedestrian-vehicle conflict helps designers to proactively implement warrants to reduce the risk at such crossings. This study uses pedestrian safety margin (PSM), a surrogate safety measure, to further define scenarios of conflict and severity. The contributing factors for occurrence of conflict were modeled and analyzed through binary logistic regression. In addition, an ordinal logit model was also developed to examine their influence on probability of occurrence of 4 different levels of severity of conflict whose thresholds were defined on the basis of Pedestrian Vehicle Scaled Risk Indicator (PVSRI). Results show remarkable goodness of fit for conflict model (AUC=91%, A=84%) and ordinal severity models (A=57.3%). Pedestrian speed, waiting time, vehicle type, pedestrian group size, accepted vehicular gap size, nature of crossing and lane position were found to have significantly impacted the odds of conflict and higher severity. The model results were further illustrated with the help of partial dependence plots and simulation plots, wherever applicable, mainly, to ascertain the interaction effects of variables.

# ACKNOWLEDGEMENT

I would like to express my sincere gratitude to Department of Civil Engineering, Institute of Engineering, Pulchowk Campus for giving me a valuable opportunity to carry out this thesis work. I am also immensely grateful to my supervisor Asst. Prof. Dr.Pradeep Kumar Shrestha for his continuous assistance throughout the period.I must express my gratitude to Asst. Prof. Anil Marsani, Program Coordinator, M.Sc. Program in Transportation Engineering and Asst. Prof. Dr. Rojee Pradhananga for their guidance. Last but not least, I am indebted to my family, friends and colleagues for their support and encouragement.

Name: Ashish Banstola
RollNo.: 076/MsTRE/003

# TABLE OF CONTENTS

**COPYRIGHT** ..... **1**
**ABSTRACT** ..... **3**
**ACKNOWLEDGEMENT** ..... **4**
**TABLE OF CONTENTS** ..... **5**
**LIST OF TABLES** ..... **7**
**LIST OF FIGURES** ..... **8**
**LIST OF ACRONYMS AND ABBREVIATIONS** ..... **10**
**CHAPTER 1 : INTRODUCTION** ..... **11**
1.1 Background ..... 11
1.2 Problem Statement ..... 14
1.3 Objective of Study ..... 14
1.4 Scope of Study ..... 14
1.5 Limitations of Study ..... 15
1.6 Organization of Report ..... 15
**CHAPTER 2 : LITERATURE REVIEW** ..... **17**
2.1 Road Crossing Behavior of Pedestrians ..... 17
2.2 Pedestrian Gap Acceptance Models ..... 17
2.3 Safety Margin and Pedestrian-Vehicle Conflict Models ..... 18
2.4 Other Risk Analysis Methods ..... 21
2.5 Summary of Literature Reviews ..... 22
**CHAPTER 3 : METHODOLOGY** ..... **24**
3.1 Research Design ..... 24
3.2 Study Area ..... 25
3.2.1 Baneshwor Midblock crossing ..... 25
3.2.2 Ekantakuna Midblock crossing ..... 26
3.2.3 Dhumbarahi Midblock crossing ..... 27
3.2.4 Thasikhel Midblock crossing ..... 28
3.3 Concept of Model Development ..... 29
3.4 Variables Definition ..... 30
3.5 Model Framework. ..... 36

3.5.1 Binary Logit Model for PV conflict. ..... 36
3.5.2 Ordinal logit model for severity of P-V conflict. ..... 37
3.5.3 Model Validation ..... 38
3.6 Sensitivity Analysis ..... 41
3.7 Sample Size Determination ..... 41
3.8 Data Collection ..... 42
3.9 Data Extraction ..... 43
3.10 Data Cleansing ..... 45
**CHAPTER 4 : RESULTS AND DISCUSSION ..... 46**
4.1 Correlation Matrix ..... 46
4.2 Results on Pedestrian -Vehicle Conflict (PVC) Model ..... 48
4.3 Kruskal-Wallis Test on Severity Levels ..... 53
4.4 Results on Pedestrian -Vehicle Conflict Severity Model ..... 53
4.5 Final Model Validation ..... 58
4.6 Sensitivity analysis through Partial Dependence Plots (PDP) ..... 61
**CHAPTER 5 : CONCLUSION AND RECOMMENDATION ..... 78**
5.1 Conclusion ..... 78
5.2 Recommendation ..... 79

# LIST OF TABLES

Table 2.1: List of Relevant Literatures ..... 23
Table 3.1:-General Confusion Matrix ..... 40
Table 3.2:-Data Information ..... 44
Table 4.1: Correlation Matrix for numeric independent variables ..... 47
Table 4.2: Cramers V Correlation Matrix for categorical independent variables ..... 48
Table 4.3: PV Conflict Preliminary Model I summary: ..... 49
Table 4.4 : PV Conflict Final Models Summary ..... 52
Table 4.5 : Kruskall-Wallis Test Results ..... 53
Table 4.6 : Preliminary Severity Model I Summary ..... 54
Table 4.7: Severity Models Summary ..... 56
Table 4.8: General Properties PV Conflict Models ..... 58
Table 4.9: Confusion Matrix of Model (2) ..... 58
Table 4.10: Properties of Model (3) ..... 60
Table 4.11: Confusion Matrix for Severity Model (3) ..... 60

# LIST OF FIGURES

Figure 3.1: Research Framework ..... 24
Figure 3.2: Baneshwor midblock crossing ..... 26
Figure 3.3: Ekantakuna midblock crossing ..... 27
Figure 3.4: Dhumbarahi midblock crossing ..... 28
Figure 3.5: Thasikhel midblock crossing ..... 29
Figure 3.6: Model Concept ..... 30
Figure 3.7: Moment Pedestrian arrives at the beginning of lane 2-Timestamp T1 ..... 33
Figure 3.8: Moment Pedestrian reaches the end of lane 2-Timestamp T2 ..... 33
Figure 3.9: Moment Vehicle reaches at the crossing in mid of lane 2-Timestamp T3 ..... 33
Figure 4.1: Correlation map for numeric datatypes ..... 46
Figure 4.2: Correlation map for categorical datatypes ..... 47
Figure 4.3: ROC curve for PVC model (2) ..... 59
Figure 4.4: PDP for Probability of conflict, Vehicular gap and Pedestrian Speed ..... 61
Figure 4.5: PDP for Probability of conflict, Pedestrian Speed and Waiting Time ..... 62
Figure 4.6: PDP for Probability of conflict vs Accepted Vehicular Gap and Group Size. 63
Figure 4.7: PDP for Probability of conflict vs Wait Time and Group Size. ..... 64
Figure 4.8: PDP for Probability of conflict vs Lane and Pedestrian Speed. ..... 65
Figure 4.9: PDP for Probability of conflict vs Lane and Wait Time. ..... 66
Figure 4.10: PDP for Probability of conflict vs Lane and Vehicle gap. ..... 67
Figure 4.11: PDP for Probability of conflict vs Wait_Time. ..... 68
Figure 4.12: PDP for Probability of conflict vs LE_ILLE ..... 69
Figure 4.13: Simulated boxplot for Probability of conflict vs LE_ILLE. ..... 69
Figure 4.14: Simulated Probability Density Plot for Probability of conflict vs LE_ILLE. ..... 70
Figure 4.15: PDP for Probability of conflict vs LE_ILLE and Wait_Time. ..... 70
Figure 4.16: PDP for Probability of conflict vs Pedestrian Speed and Wait_Time. ..... 71
Figure 4.17: PDP for Probability of conflict vs Age groups and Wait_Time. ..... 72
Figure 4.18: PDP for Probability of conflict vs Lane, Vehicle Type and LE_ILLE. ..... 72
Figure 4.19: Simulated boxplot for Probability of conflict vs Lane, Vehicle Type and LE_ILLE ..... 73

Figure 4.20: Simulated Plot for Mean Probability of conflict vs Lane, Vehicle Type and LE_ILLE ..... 73
Figure 4.21: Simulated plot for Probability of conflict vs Vehicular gap, LE_ILLE and Vehicle type. ..... 74
Figure 4.22: Simulated best fit plot for Probability of conflict vs Vehicular gap, LE_ILLE and Vehicle type. ..... 75
Figure 4.23: PDP for Probability of conflict vs Vehicle gap and LE_ILLE ..... 75
Figure 4.24: Simulated plot for Probability of conflict vs Wait Time and lane. ..... 76
Figure 4.25: Simulated plot for Probability of conflict vs Age and Vehicle gap. ..... 76
Figure 4.26: Simulated plot for Probability of conflict vs Wait Time and Vehicle type .. 77

# LIST OF ACRONYMS AND ABBREVIATIONS

SM :- Safety Margin

PSM:-Pedestrian Safety Margin

RI:-Risk Indicator

WHO:-World Health Organization

PET:-Post Encroachment Time

LPET:-Lane based Post Encroachment Time

OP:-Ordinal Probit

P-V:-Pedestrian-Vehicle

P-VC:-Pedestrian Vehicle Conflict

ROC:-Receiver Operating Characteristic

AUC:-Area Under the Curve

PDP:- Partial Dependence Plot

LE_ILLE:- Legal or Illegal

Veh_typ:- Vehicle type

Veh_gap:- Vehicle gap

PVSRI:- Pedestrian Vehicle Scaled Risk Indicator

# CHAPTER 1 : INTRODUCTION

## 1.1 Background

Midblock pedestrian crossings are considered to be a critical component of transport facility as it involves direct interaction between vehicles and pedestrians with risk of failed crossing. The interaction could be lessened by the use of signals or other traffic calming measures, but it is impractical to fully control each and every such crossing as in major urban intersections managed with signal controls, subways or footover bridges (FOBs). In addition, heterogeneous traffic and low yielding rate of drivers in developing countries has made pedestrians even more vulnerable.

Placement of bridges or subways could be an excellent option for safe crossing at midblocks but may simply not be feasible due to economic reasons and due to the danger of pedestrians crossing under/over such facilities, especially in developing countries where people often are inclined to take more risk. According to a report prepared by Ministry of Urban Development, India, most pedestrians use unprotected midblock crosswalk locations (refuge median openings with barriers, with or without zebra markings, unsignalized, and with no sign boards to control motorized vehicle drivers) because of the ease of access from their origin to their destination(MoUD India,2008). The increase in use of such midblock crosswalks significantly increases pedestrian–vehicle conflicts, and this increase further increases pedestrian fatalities at midblock crosswalks (MoUD India, 2008).

A large proportion of pedestrian deaths are reported in urban areas in low and middle income countries (WHO, 2018). In 2016, vulnerable road users (pedestrians, cyclists, and motorcyclists) accounted for approximately 72 percent of all road fatality victims, among the highest rates in the country, with pedestrians accounting for half of those (World Bank, 2020). A nationwide study conducted by Kumar and Suvash (2010) on injuries and violence found road traffic injuries as the most common injury type and nearly half (48.6%) of such injuries were borne by pedestrians. An article by Ojha (2021) studied road safety status in developing countries and found reckless pedestrian crossing as a

important cause of road accident. And since the sidewalks in urban areas are relatively safer than crosswalks, it can be reasonably concluded that such injuries occur as a result of unsuccessful crossing. In, Nepal, a total of 2,485 individuals lost their lives in road traffic crashes during FY 17/18 ; out of this, 28% were pedestrian, while 44% of pedestrian fatalities involved buses or trucks (Draft Nepal Road Safety Action Plan, 2021-2030) ,also statistic reveal 19,998 pedestrians were enrolled for a haphazard crossing pedestrian awareness class in 2020 (Ojha,2021).

Sustainable goals have envisioned pollution free and pedestrian friendly urban settings and as such, short trips on foot should be preferred compared to taking bus or taxi which can be achieved only through well designed pedestrian facilities (Patel, et al., 2018). Past researches and findings in Nepal highlighted that pedestrians were most vulnerable groups in road accidents because pedestrian safety had not been considered in design of transport system. Growing urbanization of towns, rapid expansion of road networks and increasing number of vehicle ownership across the country is set to make the situation even worse in coming days if steps are not taken towards making crossings safer for vulnerable road users. The Government of Nepal, in line with UN Global Action Plan for road safety, had introduced Road Safety Action Plan in 2013 envisioning the roles of various government bodies in reducing road crashes which amongst others includes proposition of activities and research work ensuring pedestrian safety at crossings (MoPIT Nepal, 2013).

Midblock crossings are critical junctures in terms of pedestrians safety as it concentrates space and time of multiple interaction of road users while one of the users is more vulnerable than the other. Severe conflict occur when road users fail to predict and react to other users' decisions. Further, varying behaviours of drivers and pedestrians can also lead to misunderstanding, which results into conflicts with varying severity (Chaudhari et.al, 2019).

The safety of pedestrians at crossings can be evaluated either by the use of historic crash data or through non crash measurements. Complete crash data could provide sound measurement of safety but in many cases, they are either in inadequate form or not available at all. Furthermore they only occur very few times and cannot be used to measure potential risk at a crossing where no such incidents has occurred. Another form

of measuring risk is through non-crash, often called proactive safety measurement, where surrogate measures of safety (SMOS) variables namely, Time to collision (TTC) and Post Encroachment Time(PET) are observed. It involves identifying near-miss events (narrowly escaped collisions) and seeks the actual information about the events with driver as well as pedestrian behaviour under site conditions (Kadalii & Vedagiri,2016).

Time-to-collision(TTC) is the time period in which the road users would collide with each other if they had continued at the same speed and collision course(direction). Zhang et. al, (2017) and Dhamaniya et. al, (2019) have used this technique for identifying aggressive behaviour of pedestrian and vehicle during their interaction. Post Encroachment Time(PET), the time difference between the arrival of a later road user at a potential conflict point and arrival of earlier user, is a measure of how fast the conflict point is occupied by a later road user ; less PET shows conflict would have occurred in either of these two conditions, if earlier road user arrived a bit later or the later arrived little earlier, while high PET shows safer scenario of the interaction. PET has been modified in many ways in later years to protray more real picture for pedestrian vehicle interaction for eg. Zhang et.al has evaluated pedestrian safety using lane based PET (LPET) (PET for each and every lane) and built a ordered probit (OP) model considering traffic volume, vehicle speed, pedestrian crossing behaviour, availability of refuge and so on. Similarly, Govinda et.al (2022) have proposed a new modified indicator (RI) as the ratio of vehicle speed to PET in their paper to quantify pedestrians' risk. Recent literatures have been found using a more specific term, safety margin(SM) often instead of PET.

Conflicts and crashes are random events whose occurrences are influenced by various external factors such as traffic control, geometric design (Zhang, et al. 2013) , behavioral and vehicular characteristics of the road users and so on,. If we could assess the influence of these factors on the probability of conflict occurrence then it would be helpful for transport planners in advance to anticipate the risk so as to proactively change the design or policy for safer crossings. This research is an attempt towards that.

Risk analysis is also necessary to develop cost-effective countermeasures capable of reducing the risk of pedestrian crash. Quantification of relationship between pedestrian and vehicle characteristics, site characteristics and risk/conflicting behavior of road users can help us to identify the role of crucial variables in order to apply measures in the

direction of safer crossings which may be achieved through change in pedestrian behavior, road design, speed reduction and so on.

## 1.2 Problem Statement

Sustainable transport demands walking be made a viable mode choice especially in built-up urban environments. As such, safer crossings become an integral part of sustainable urban transport system. Unlike in developed countries, due to mixed traffic, absence of signal control and pedestrian facilities and low yielding and non lane based driving, the pedestrian vehicle interaction in growing cities is more complicated and risky. Low-income countries continue to report increases in traffic fatalities and the proportion of pedestrian fatalities have remained high (WHO, 2018).Thus, there is a need to evaluate the safety of existing midblock crossings proactively and identify significant risk contributing factors.

## 1.3 Objective of Study

The main objective of this study is to evaluate pedestrian safety at uncontrolled midblock crossings. Specific sub-objectives are:

1.To develop a Pedestrian-Vehicle conflict model to quantify the effects of pedestrian behavior, demographics, vehicular and road characteristics on occurrence of conflict.

2.To formulate a Pedestrian-Vehicle conflict severity model to understand the influence of aforementioned factors on severity of such conflict.

## 1.4 Scope of Study

The study is carried out at four uncontrolled midblock crossings inside Kathmandu valley. Following are some major scope of the work:

1.To identify and measure various factors affecting the safety of pedestrian (safety margin) at uncontrolled midblock crossings.
2.To analyze the magnitude of influence of those factors on Safety Margin.

3.To conduct a pedestrian - vehicle conflict analysis based upon the safety margin values and the factors associated with it.
4.To suggest measures that could decrease the potential risk for pedestrians during crossing.

## 1.5 Limitations of Study

Limitations of the study are listed as follows:

- The study is limited to four lane undivided midblock crossings.
- Vehicle flow rate, pedestrian flow rate were not taken into account.
- Minute pedestrian behaviour like hand signals, frequency of attempt, looking at phones were ignored.
- Land use type was also not taken into account.
- Pedestrian/vehicle trip purpose could-not be assessed.
- Pedestrian demographics/behaviour, vehicular characteristics and were observed from videographic survey.
- Conflict was defined on the basis of previous literatures, an actual perceived conflict couldnot be taken into account.

## 1.6 Organization of Report

This report consists of following five chapters:

**Chapter 1: Introduction**- provides the background of the study, introduces pedestrian safety and risk, also defines the problem statement, objective and scope of the research and its limitations.
**Chapter 2: Literature Review**- comprises of various literatures mostly related to safety margin, discusses other risk analysis methods and their approaches/considerations..
**Chapter 3: Research Methodology**- includes the description of site, variables; outlines the model framework and the overall flow of the study.
**Chapter 4: Results and Discussion**- describes the output of the model, interpretation of the model coefficients and illustrations through partial dependence plots and simulation plots.

**Chapter 5: Conclusions and Recommendation-**sums up the study with major conclusions and provides recommendation for designers and researchers for further study.

# CHAPTER 2 : LITERATURE REVIEW

## 2.1 Road Crossing Behavior of Pedestrians

Road crossing behaviour of pedestrians is in inherently tied with their safety. Extensive research have been done on crossing behavior of pedestrians with factors like pedestrian perception, roadway and environmental characteristics taken into account. Earlier studies provide significant facts about pedestrian demographic characteristics (such as age, gender) and how these characteristics influence road crossing behaviour. Such studies have focused on detailed experiments to find out the effect of age on road crossing decisions with effect of vehicle distance or speed of vehicle (Oxley et al., 1997; Lobjois and Cavallo, 2007). Road crossing behaviour with respect to gender has also been observed in various studies. Males have a tendency to show more hazardous road crossing behaviour than females due to less waiting time (Khan et al., 1999; Tiwari et al., 2007). Few studies have also explored the importance of the pedestrian speed at different locations (Knoblauch et al., 1996), such as the zebra crossing location (Varhelyi, 1998) and signalized intersections (Tarawneh, 2001). Outline of these studies suggest that males walk significantly faster than females while crossing the roads. A recent study was focused on legal versus illegal pedestrian road crossing behaviour at mid-block location in China (Cherry et al., 2012). Few studies have identified pedestrian behaviour in mixed traffic streets and developed a microsimulation model in order to find out the fundamental characteristics as well as the conflicts of the pedestrian movement (Shahin, 2006).

Some studies have also addressed pedestrian road crossing behaviour by considering the effectiveness of educational training programs (Dommes et al., 2012). Studies had identified the importance of the environmental characteristics, such as type of crossing facility, traffic volume and roadway geometry on road crossing behaviour. Some studies have also explored the pedestrian road crossing behaviour before and after re-construction of traffic facility (Gupta et al., 2010).

## 2.2 Pedestrian Gap Acceptance Models

Pedestrian Gap acceptance models considering internal and external factors have been used to assess the safety of pedestrians at unsignalized midblock crossings for a long time. These mathematical models are often discrete choice models (mostly binary logit) quantifying the relationship between probability of accepting certain gap with various factors like pedestrian demographic, waiting time, rolling behaviour ,vehicle speed and gap size, pedestrian speed, road width, traffic volume and so on.

Yannis, et al.(2013) carried out the study in Athens at an uncontrolled midblock location. Videographic survey was done in real traffic conditions and a binary logit model was built with independent variables namely, vehicle type, waiting time, presence of parking and vehicle gap which were found significant to the model. Waiting time and presence of parking were negatively correlated with crossing decision while presence of large traffic gap increased the probability to cross. Zhao, et al.(2019) conducted a similar study which showed that gap size and crossing distance have highest effect on the gap acceptance decision and further, higher waiting time showed more probability to accept dangerous gaps for which the research recommends the installation of "YIELD" sign to remind the driver that pedestrians may cross such condition. Similarly it was found that crossing distance of more than 12m was more risky and needed median islands.

### 2.3 Safety Margin and Pedestrian-Vehicle Conflict Models

The concept of safety margin was first found to be introduced by Oxley et al.(1997) where study was done on differences in behavior of older and younger pedestrians while crossing a two lane undivided road to find out if the decline of cognitive, physical, sensory and perceptual abilities in older generation increases their vulnerability while crossing. The result showed that younger slow walkers left a larger safety margin while the older slow walkers placed themselves at an increased risk of collision by keeping very less safety margin approving the hypothesis.

Chaudhari et al. (2019) built a Multiple Linear Regression model (MLR) to comprehend the factors influencing pedestrian safety with safety margin as a dependent variable. Videographic survey at four different uncontrolled midblock crossings in Indian cities was done from which eleven different variables under pedestrian demographics, vehicle characteristics and roadway characteristics were extracted. The model illustrated negative

correlation of safety margin with rolling behaviour, presence of light vehicles and platoon size, while safety margin increased with increase in vehicle gap, age and pedestrian speed.

Kadali and Vedagiri(2015) conducted similar study at eight different midblock locations in India. In addition to MLR model, a binary logit model for Pedestrian vehicle non-conflict (PVNC) prediction was constructed to see the factors influencing probability of pedestrian-vehicle conflict. Increase in age showed decrease in SM and thus, increase in probability of conflict (PC) while gender was not found significant to the model. Under pedestrian behavioral characteristics, rolling behavior was negatively correlated and platoon size was positively correlated with both SM and PC. Increase in accepted vehicular gap size also showed increase in SM. The results illustrated that pedestrians took more risk with two wheelers and 3 wheelers with less safety margin. Also, pedestrians crossed with more safety as the no of lanes (road width) increased .

Kadali and Vedagiri (2016) assessed the severity of pedestrian vehicle conflict at unprotected midblock crossings in India. An ordered probit (OP) model was built in which the dependent variable was levels of severity of conflict based upon the distribution of safety margin values. Eight different midblock locations with varying land use and road characteristics were selected for study.The OP analysis revealed that pedestrian behavioral characteristics like rolling behavior and pedestrian speed change condition increase the severity of conflict. On the contrary, more waiting time was associated with less severe conflict. Further, it was shown that lack of traffic barrier (median) increased PCS levels. Sites with mixed land use compared with residential,school and commercial showed higher levels of PCS.

A more detail approach was taken by Zhang et al.(2017) where, unlike aforementioned studies safety margin was observed for each lanes. They conducted video graphic survey at five multi lane crosswalks in Wuhan city of China. Number of conflicts was taken as an ordinal dependent variable influenced by traffic volume, presence of refuge, crossing strategy among others. OP model analysis showed percentage of conflict for female less than for males but percentage of serious conflict was greater for females than for males. Moreover, the study showed rolling crossing more dangerous in multilane crosswalks and also, increase in traffic volume, speed and absence of pedestrian refuge contributed to

higher conflicts. Based on their field observation and model, the number of conflicts will rise by 2% while the traffic volume increases 200 pcu/h; similarly, if the vehicle speed increases 5 km/h, the number of conflicts will rise by 12% accordingly.

Chaudhari et al.(2020) developed a binary logistic model to predict probability of pedestrian avoiding conflict with approaching vehicle. Model showed that pedestrian's decision to cross the road with or without safety depends upon vehicle type and speed, pedestrian speed, vehicular gap available, number of lanes, pedestrian rolling behaviour, land use, pedestrian age, platoon size, accepted gap/lag, type of gap. Further, the increase in the vehicle speed resulted in the increase PVNC. Pedestrian's individual characteristics were found to be insignificant whereas gender was found to have an impact on SM. Rolling behaviour, presence of marking, number of lanes discourages pedestrians towards the decision to cross, regardless of the traffic gap.

Similar study was done in Korea to evaluate pedestrians' risk at unsignalized crossings. Lee and Jang (2018) found older people were at greater risk than the young ones. There was an insignificant difference between the SM of approaching vehicles that were traveling at speeds less than 30 km/h and those traveling at speeds in the range of 30-50 km/h. Further, when the speed of the vehicles exceeded 50 km/h, the risk of conflict was higher than it was for vehicles traveling at speeds below 30km/h. The ratio of conflict risk for crossing gradient topography road was found to be about 21.7 times greater than that for the non-gradient topography area. Regarding safety facilities, the 30 km/h speed limit sign influenced the risk situation of conflict. The ratio of conflict risk for a road with the safety facility was found to be about 0.395 times lower than that for an unmarked road. Interestingly, the effect of a marked crosswalk without a traffic signal was insignificant, so the result showed that a marked road without a signal has no effect on safe crossings.

Almodfer et al. (2015) studied lane based distribution of severity of conflicts in 4 lane divided crosswalks. Results showed that conflicts were not distributed evenly over the four lanes. For each traffic flow direction in the study site, the far lanes recorded a higher percentage of serious conflicts than the near lanes, and slight conflicts were the most frequently occurring conflicts for both directions. Analytical results showed that shorter waiting time (less than 3 s) caused 881 conflict situations between pedestrians and vehicles. As pedestrian waiting time went from 3 s to 30 s, serious conflicts decreased

significantly (from 194 to 9). When pedestrian waiting time went beyond 30 s, pedestrians crossed in very risky situations.

Cherry et al.(2012) observed illegal midblock crossing in six lane road of Wuhan, China. Conflict was categorized into two group namely, speed change conflict and lane change conflict; three different probit models were built to predict the conflict of first type, second type and conflict occuring due to any one of those. The results showed effective gap, vehicle speed and crossing step as significant variables, all negatively correlated with the conflict probability. Moreover, a bivariate conflict and gap acceptance model was built which showed results of similar nature; in addition, here, near lanes had more chance of conflict after gap acceptance.

Zhuang et al.(2010) explored pedestrian safety at unmarked crossings in China. Stepwise regression model was built which revealed following results. Pedestrians' higher looking frequency rather than longer duration at vehicles before crossing could improve their safety, but looking behaviors during crossing did not play an important role. Higher frequency of looking before crossing was found to increase safety with more safe behaviors. Increase in crossing speed, group size, going backwards and looking frequency increased the safety margin values and thus pedestrians safety.

Govinda et.al(2022) argued that PET cannot solely define the risk and used a new indicator(Risk Indicator (RI)), the ratio of approaching vehicle speed and PET. Videographic survey was done at 4 legged uncontrolled intersections from which pedestrian and vehicular data was extracted ; a MLR model was built considering RI as independent variable and pedestrian speed, gender, vehicle type as outcome variables. Later, using Support Vector Machine algorithm(SVM), threshold values of RI for pedestrian speed, gender and vehicle type was also generated. MLR results showed that pedestrian gender, age and speed, vehicle type and speed, interaction location and crossing position have a significant effect on RI.

### 2.4 Other Risk Analysis Methods

Historic crash databases are also often used to identify pedestrian crash patterns. Studies generally point to male pedestrians as those most frequently involved in pedestrian

crashes, and they point to the elderly and children as the most vulnerable pedestrians .Nowadays they are ousted by conflict based techniques because of their unreliability and inadequateness in many cases. For eg. Diogenes and Lindau (2010) used crash based technique to evaluate safety at midblock crossings in Brazil.The study demonstrated that crossings located close to bus stops, or busway systems, experienced higher pedestrian crash rates.

Acharya and Marsani (2019) studied the relationship of illegal midblock crossing volume with traffic and geometric parameters of road. A MLR model was built which showed that volume of such illegal crossings depends only upon traffic speed and other factors like traffic volume, presence of crossings nearby and carriageway width are irrelevant.

Devkota and Shahi (2013) specified and estimated proportional hazard models to identify the determinants of pedestrians waiting time (delay) on pedestrian crossings. According to the study, male pedestrians, pedestrians with group, pedestrians going to work and well educated pedestrians were likely to accept higher risk and cease their waiting time at pedestrian crossings.

Pedestrians past involvement or witness to accidents inhibited longer waiting time before his/her successful crossings. Pedestrians in group were more likely have lesser waiting time than one crossing individually. Furthermore, pedestrians who had access to private vehicles seemed to be more aware of risk involved. Therefore, they were more cautious of time needed before crossing (Devkota and Shahi, 2013).

### 2.5 Summary of Literature Reviews

Studies have tried to discern the risk taking behaviour of pedestrians by observing mainly their demographics, crossing behaviour and road & traffic characteristics.In most of those studies conflict is categorized as a binary variable and logistic or probit regresion has been performed while in some, ordered probit model has been built considering levels of conflict as an ordinal variable. Literatures ( for eg. Kadali & Vedagiri, 2016) have considered SM as the time difference between pedestrian reaching the median/curb and vehicle arriving at the potential conflict zone; this approach often misleads the interpretation of safety margin because according to this definition, in higher lane roads,

negative safety margin is more likely to happen, which doesnot necessarily reflect severe conflict conditions. It is more accurate to calculate SM on lane by lane basis (for eg in. Zhang, et al. 2017) - termed as Lane based Post encroachment time (LPET).

In most of these literatures, conflict has been defined as the scenario where a pedestrian crosses the road with safety margin value less than 1 sec. Some have further divided the conflict into different ordered levels ; Zhang, et al (2017) considered three conflict levels based on the LPET values; serious conflict( <1 sec), slight conflict(1-3 sec) and No conflict (>3 sec) whereas Kadali & Vedagiri (2016) divided conflict into 6 levels based on the cumulative distribution of SM values. The minimum threshold for conflict can be found 0.7sec in some literatures.

Waiting time, rolling behavior, vehicle gap, vehicle speed and pedestrian speed were regarded as some important factors in most of these literatures. Road characteristics like number of lanes and presence of median or any traffic facility also played a crucial role in pedestrian's decision to cross with risk. Assessment of these influencing factors can help designers to proactively evaluate risk to pedestrians at a newly made crossing and apply warrants to increase the safety.

**Table 2.1: List of Relevant Literatures**

| Study | Model Framework | Dependent V | Independent V |
|---|---|---|---|
| Chaudhari et al. (2019) | MLR | SM | Pedestrian behaviour,demographics,vehicle and road characteristics,driver behaviour |
| Kadali and Vedagiri(2015) | MLR and Binary Logit | SM and PVNC | |
| Zhang et al.(2017) | Ordinal Probit | Number of Conflict | |
| Kadali and Vedagiri (2016) | Ordinal Probit | Severity of Conflict | |
| Chaudhari et al.(2020) | Binary Logit | PVNC | |
| Lee and Jang (2018) | Chi-squared analysis | SM | |
| Almodfer et al. (2015) | Analytical | Severity of Conflict | |
| Cherry et al.(2012) | Probit and Bivariate Probit | Conflict | |
| Zhuang et al.(2010) | MLR | SM | |
| Govinda et.al (2022) | MLR | Severity of Conflict | |

# CHAPTER 3 : METHODOLOGY

## 3.1 Research Design

This research aims at predicting the probability of pedestrian-vehicle conflict at unsignalized midblock crossings and analyzing various factors influencing it. Relevant literatures were reviewed from which variables to be associated in the model, type of model etc were identified. A framework, as depicted in Figure 3.1, was then created for guiding the flow of the study. Suitable sites were identified where video graphic survey was done to observe the pedestrian-vehicle interaction and extract necessary variables describing pedestrian characteristics, road characteristics, and vehicle characteristics.

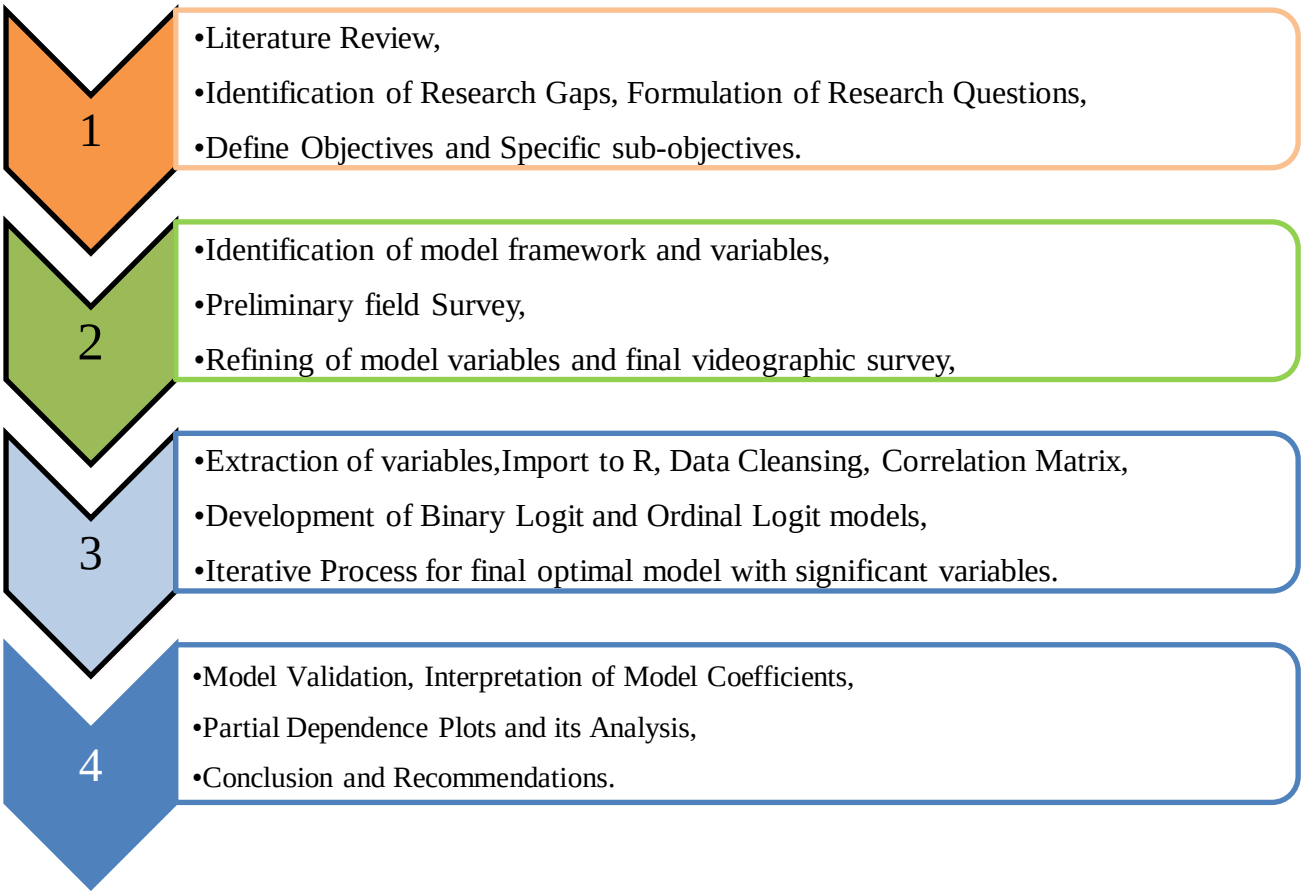


**Figure 3.1: Research Framework**

### 3.2 Study Area

For the proposed study, ideal site would be an uncontrolled unsignalized midblock crossing (legal/illegal) where pedestrians cross perpendicular to the direction of movement of vehicles with no or very little amount of side friction. The factors influencing pedestrian-vehicle interaction like crossing decision, increasing/decreasing speed, how much time to wait in curb, whether to yield vehicle or not, should not be guided by external control factors like YIELD, STOP signs, or traffic signals. Furthermore, the crossings should be sufficiently away, preferably more than 100 meters, from intersections to avoid the effect cautious deceleration/acceleration/turning of vehicles (hindered).Also the side friction should be minimum meaning that:

1.The vehicle entering/exiting from/to sideways, U-turning in the main road section shall be minimal
2.The presence of parking would affect the speed of vehicle and therefore would interfere in the pedestrian vehicle interaction so , no parking near curb.

Also the section should be straight with no any impedance (like potholes or other pavement defects) to running vehicle. Higher pedestrian flow would be preferred for gathering large amount of data.

Based on the above criteria four midblock crossings, three located at ring road sections and one at main road of Baneshwor were selected whose characteristics are explained in detail below.

#### 3.2.1 Baneshwor Midblock crossing

This section is located in front of Everest Hospital at Baneshwor, about 265 m west of the main intersection. The road cross section can be divided into two parts. The service lanes at north side and the south side (part one) are specially designed for public vehicles to pick and drop passenger along the curb. Deceleration, acceleration and stoppage of these vehicles affected the speed of other following vehicles and also disrupted ideal pedestrian cross flow scenario .On contrary, the other part, main road which consists of

four lanes-two lanes for each direction, was found to have no such side friction obstructions and met all other basic site selection criteria. Thus, observation was initiated/terminated once pedestrian reached one of the two shelters (median like structure separating the two parts).

**Figure 3.2: Baneshwor midblock crossing**

The salient features of this crossing are:-

Uncontrolled and unsignalized
Lane width:-3.01m
Total Length of crossing :- 12.41m
Width of crossing:-4.4m
Availability of Pedestrian shelter at both crossing ends
Presence of lane markings:-Yes
Legal/Illegal Crossings:-Legal (presence of zebra marking)

### 3.2.2 Ekantakuna Midblock crossing

This section is located in front of Department of Transport Management Office at Ekantakuna in ring road. Similar to the Baneshwor crossing, this road section also can be divided into two parts- main and service road. The observations from the service road was found to be affected by side frictions like parking, public vehicle loading and unloading, vehicle entering/exiting, from/to service lanes and therefore, omitted whereas, the main road- inner middle four lanes, free from all these obstructions, was included for the study.

Unlike Baneshwor section, here, passenger shelter is located only on the north side of the road crossing; on the south side, there is a concrete block which pedestrian mount on to enter/exit the main road.

**Figure 3.3: Ekantakuna midblock crossing**

The salient features of this crossing are:-

Uncontrolled and unsignalized
Lane width:-3.5m
Total Length of crossing:-14.32
Availability of Pedestrian shelter at only one end
Presence of lane markings:-Yes
Legal/Illegal Crossings:-Illegal (absence of zebra marking)

### 3.2.3 Dhumbarahi Midblock crossing

Located at the northern ring-road section near Gopi Krishna Bridge in Kathmandu, this uncontrolled midblock crossing comprises of four undivided lanes. Lane markings were found but were not distinct. Although, ideal conditions for observation were not met at some instances when public vehicles dwelled or vehicles crossed the road, these scenarios were not substantial, allowing to consider the pedestrian-vehicle interaction while ignoring such cases. In contrast to two crossings earlier, here, the whole road width is of four lanes without any physical structure or passenger shelters in between.

The salient features of this crossing are:-

Uncontrolled and unsignalized
Lane width:-3.3m (Average)- varies across lanes
Total Length of crossing:-13.2m
Presence of lane markings:-Yes
Legal/Illegal Crossings:-Illegal (absence of zebra marking)
Presence of minor side friction due to dwelling of public vehicles.

**Figure 3.4: Dhumbarahi midblock crossing**

### 3.2.4 Thasikhel Midblock crossing

This is a legal midblock crossing, at a section in southern ring road. The road characteristics of this crossing in similar to the Ekantakuna crossing except the fact that this one is legal. Other salient features are listed below:

The salient features of this crossing are:-

Uncontrolled and unsignalized
Lane width:-3.2m (Average)- varies across lanes
Total Length of crossing:-14.8m

Width of crossing:-3.65m

Presence of lane markings:-Yes

Legal/Illegal Crossings:- Legal (Presence of zebra marking)

**Figure 3.5: Thasikhel midblock crossing**

### 3.3 Concept of Model Development

Attempt was made to establish relationship between various independent variables with conflict. At first, collinearity test was done to identify whether the independent variables have strong relationship in themselves which if true may create redundancy in the model. Then, following types of model were developed:

1. A binary logit Pedestrian-Vehicle conflict model to quantify the effects of pedestrian behavior, demographics, vehicular and road characteristics on occurrence of conflict.

2. An ordinal Pedestrian-Vehicle conflict severity model to understand the influence of aforementioned factors on severity of such conflict.

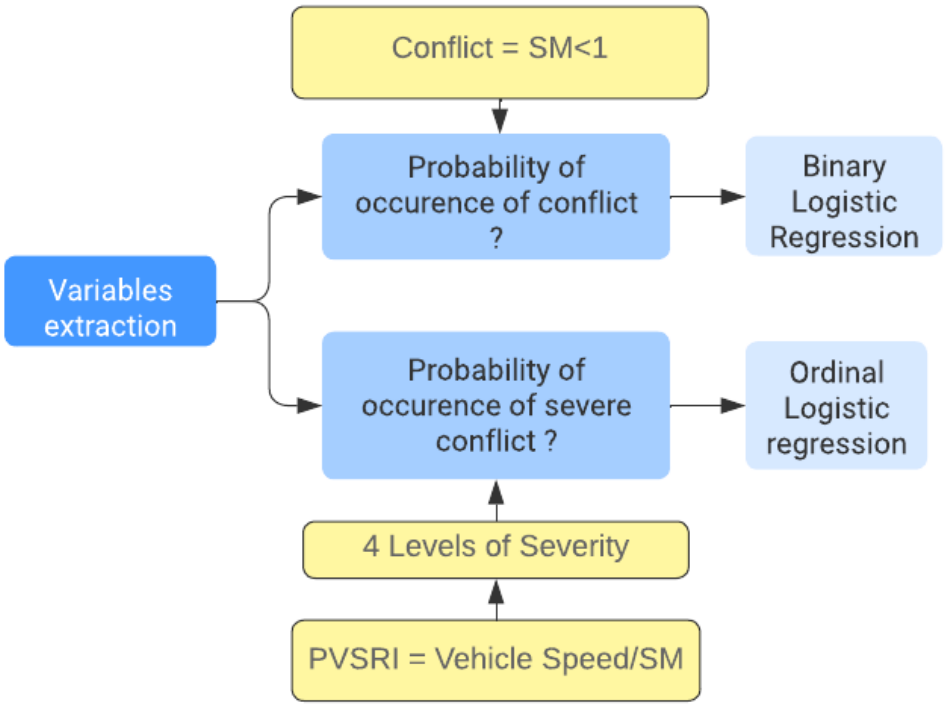


**Figure 3.6: Model Concept**

Diagrammatic representation of the model building is shown in Figure 3.6. Stepwise regression was carried out for finding appropriate combination of variables for the final model which was later validated and then interpreted accordingly.

## 3.4 Variables Definition

Various literatures have assumed some common set of independent variables influencing conflict. Pedestrian demographics, pedestrian behavior, roadway and vehicular characteristics and driver behavior are taken into account in an attempt to explain the conflicting behavior of pedestrian.

### 3.4.1 Pedestrian Demographics.

For explaining the influence of demographics, pedestrian gender and age are observed visually from the video.

**a) Gender**

Gender is assumed to be a categorical variable. In the observation sheet, male is referred as "M" and female is referred as "F".

**b) Age**

Age is also assumed to be a categorical variable divided under three groups:

i)Above 40 years old:- Category 3
ii) From 20-40 years old:- Category 2
iii) Below 20 years old:- Category 1

This was extracted purely based on observation from the video. Categorical instead of ordered integer was assumed owing to the fact that the conflicting behavior may not increase/decrease with age. People in middle age category (2) may manifest high conflicting behavior than of the age groups above or below them.

### 3.4.2 Pedestrian Behavioural Characteristics

Crossing in group, waiting time, pedestrian crossing speed, accepted vehicular gap are observed to describe pedestrian behavior.

**a) Group Size**

Conflicting behaviour also depends upon the number of people that the pedestrian cross the road along with. It is assumed to be categorical, divided into four categories:

i)Single :- Category 1
ii)Two people crossing at the same time:-Category 2
iii)Three to four people crossing at the same time:-Category 3
iv)More than four people crossing at the same time:-Category 4

**b) Pedestrian Waiting Time**

Literatures suggest that the more time a pedestrian waits at the curb, the more conflicting behaviour he/she shows. Here, waiting time is taken as a continuous variable. It is measured as the time difference between the pedestrian arriving at the beginning of a lane and his/her departure from the lane after finding suitable vehicular gap.

**c) Pedestrian Speed**

Previous researches show that with increase in speed there is increased chances of conflict. Pedestrian speed is calculated on a lane by lane basis (for each lane).It is assumed to be a continuous variable taking values greater than zero.

**d) Pedestrian Accepted Vehicular Gap**

Vehicular gap refers to the temporal gap of consecutive vehicles. Pedestrians were observed to wait at the beginning of lane until they found convenient gaps to cross a lane. It is considered as a continuous variable.

### 3.4.3 Safety Margin and Conflict (Dependent variable)

Safety margin is the time difference between a pedestrian reaching the end of each lane/curb and vehicle arriving at crossing. Higher values of SM indicate non conflicting behaviour while lower values indicate conflicting behaviour of pedestrians. In this study, safety margin (independent continuous variable) is calculated on a lane by lane basis that is, if a pedestrian crossed a 4 lane road there would be four safety margin values.

The concept of safety margin is illustrated through Figures 3.7, 3.8 and 3.9 to represent a particular case of pedestrian crossing. For instance, suppose a pedestrian has crossed Lane 1 and arrives at the beginning of Lane 2 as depicted in Figure 3.7 at time T1 and after a while, on finding suitable vehicular gap, the pedestrian crosses the lane and eventually reaches at the end of Lane 2 at time T2, as shown in Figure 3.8. The next Figure 3.9 shows a two wheeler that arrives at the middle of the lane at Time T3 while the pedestrian reaches the middle of the other lane. Here, Safety Margin is the difference in time T3 and T2.

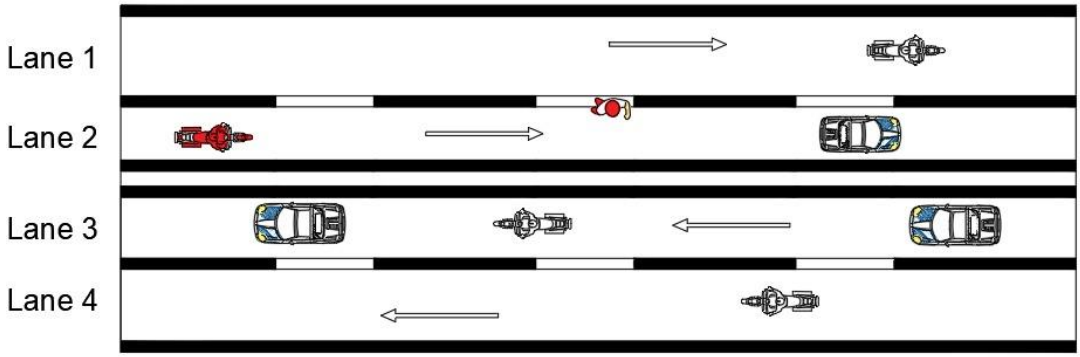


**Figure 3.7: Moment Pedestrian arrives at the beginning of lane 2-Timestamp T1**

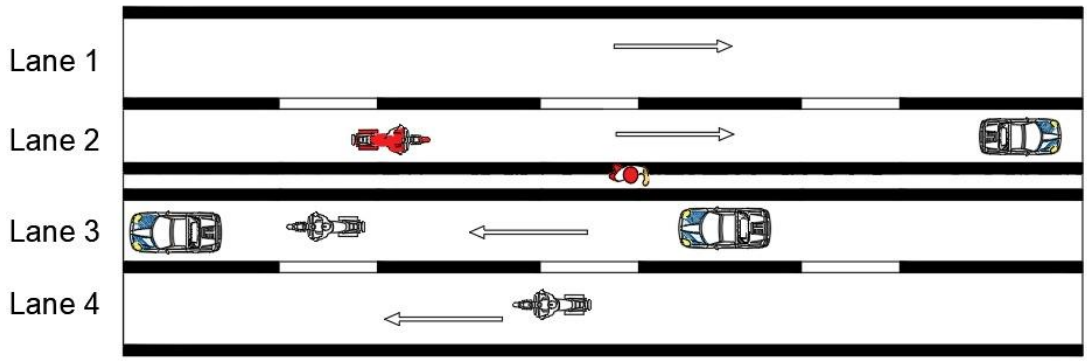


**Figure 3.8: Moment Pedestrian reaches the end of lane 2-Timestamp T2**

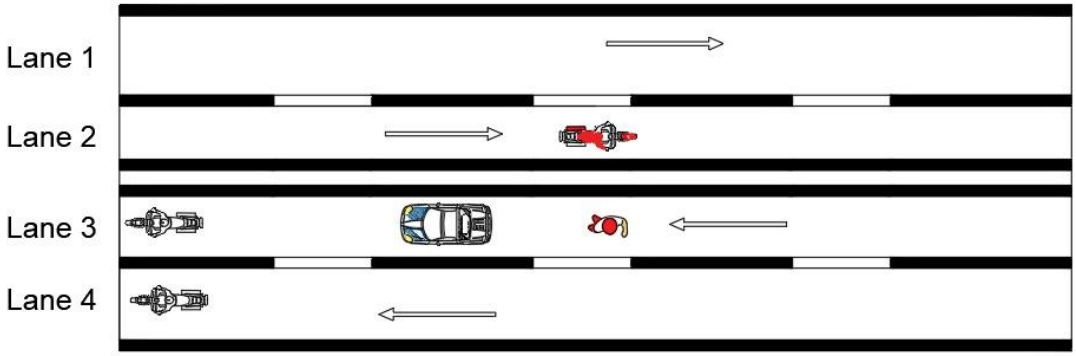


**Figure 3.9: Moment Vehicle reaches at the crossing in mid of lane 2-Timestamp T3**

Furthermore, conflict is a scenario where this temporal measure of safety, the margin value, is less than 1; so, every pedestrian crossing a lane has two outputs, either he/she shows conflicting behavior (SM less than 1) or non-conflicting behaviour (SM more than 1). Threshold of 1 sec has been used by Kadali and Vedagiri (2016), Zhang et al.(2017), Almodfer et al. (2015), Chaudhari et al.(2020) and many other literatures with the notion being that pedestrian and driver need at least 1 second for reaction time and failure to secure that reflects the inability of the pedestrian to cross without conflict.

i) Conflict :Yes (Safety margin value less than 1)(Category 1)
ii) Conflict: No(Safety margin value greater than 1) (Category 0)

### 3.4.4 Ordinal Levels of Severity of Conflict (Dependent variable)

Risk Indicator(RI), dependent variable for ordinal logistic regression is a concept which arises from the inadequacy of safety margin to define the severity of conflict because of latters relation with vehicle speed; safety margin of 2s against 50km/h vehicle speed is a severe case than against 20km/h vehicle speed. Thus, RI is commonly defined as the ratio of vehicle speed to safety margin(SM). However, since SM contained negative values there was a need to rescale it to positive values greater than zero which was done by shifting the minimum negative value of SM to zero and raising other values accordingly. The new risk indicator thus obtained has been termed as Pedestrian Vehicle Scaled Risk Indicator(PVSRI).

Mathematically,

$$PVSRI = \frac{Vehicle\ speed}{Scaled\ SM}$$

Four levels of severity has been defined based on the distribution of this PVSRI.CDF values at 0,0.25,0.5, 0.75 and 1 were considered to be the threshold values for "No Risk", "Slight Risk", "Fair Risk" and "High Risk" severity levels.

i) No Risk: RI at 0% CDF of RI to RI at 25% CDF of RI

ii) Slight Risk: RI at 25% CDF of RI to RI at 50% CDF of RI
iii) Moderate Risk: RI at 50% CDF of RI to RI at 75% CDF of RI
iv) High Risk: RI at 75% CDF of RI to RI at 100% CDF of RI

### 3.4.5 Vehicular Characteristics/Driver Behaviour

#### a) Vehicle type

Vehicle type is another factor influencing the conflicting behaviour. Researches show that for motorcycles, pedestrians tend to cross with less safety margin value (higher conflicting behaviour) while for trucks the value is much larger. Here vehicles have been divided into four categories;

i)Motorcycles/Scooter: Category "Two_Wheeler"
ii)Car: Category "Car"
iii)SUV/Buses/Van: Category "Buses_Van"
iv)Heavy trucks: Category "Heavy"

#### b) Vehicle Speed

Speed of vehicle is taken as a continuous variable. Distance between references was taken from the field and time for vehicle to cover 30m for illegal and 15 m for legal intersection was then noted, after which speed was calculated as ratio of distance to time.

### 3.3.6 Driver Yields or not

This denotes the way a vehicle is reducing speed on approaching pedestrian.It has been classified into three ordered categories (0,1 and 2).If a vehicle reduces speed abruptly , the value is 2; while 0 is for those who donot change their speed upon interacting with pedestrian. Value 1 is given for slight deceleration of vehicles.

### 3.4.7 Roadway Characteristics

Lane width, legal/illegal crossings are also observed as a part of road characteristics.

i) Lane width : It is a continuous variable assuming value greater than zero.

ii)Legal/Illegal Crossings: Whether a crossing is legal (presence of zebra crossings) or not; a binary categorical variable assuming value 0 for illegal and 1 for legal crossings.

### 3.4.1 Model Framework.

### 3.4.2 Binary Logit Model for PV conflict.

When the outcome is a binary variable, it is often preferred to portray it in the form of binary logistic model. Here, conflict has been defined as the case where the SM value is less than 1 second ; so every pedestrian crossing a lane has two outputs, either he/she shows conflicting behavior (SM less than 1) or non-conflicting behaviour (SM more than 1).Each of these two cases have their own set of independent variables (pedestrian demographics, behaviour, roadway/vehicle characteristics, and so on ).

$$\ln\left(\frac{\text{prob(conflict)}}{1-\text{prob(conflict)}}\right) = \beta_0 + \beta_1 X_1 + \beta_2 * X_2 + \ldots + \beta_k * X_k + \in \qquad \ldots (3.1)$$

where,

$\beta_0 = \text{Constant}$

$\beta_k = \text{Coefficient describing the role of variable X1 in the model}$

$X_k = \text{Value of variable } X_k \text{ in the model}$

$\in = \text{ error of the model}$

Rearranging the equation gives:

$$\text{prob(conflict)} = \frac{1}{1 + e^{(-(\beta_0 + \beta_1 X_1 + \beta_2 * X_2 + \ldots + \beta_k * X_k))}} \qquad \ldots (3.2)$$

Similarly,

$$\text{prob(no conflict)} = \frac{e^{(-(\beta_0 + \beta_1 X_1 + \beta_2 * X_2 + \ldots + \beta_k * X_k))} 1}{1 + e^{(-(\beta_0 + \beta_1 X_1 + \beta_2 * X_2 + \ldots + \beta_k * X_k))}} \qquad \ldots (3.3)$$

Equation (3.2) and (3.3) are also called sigmoid function. The determination of whether the case (set of independent variables) falls under conflict category or not further depends upon threshold value. Commonly, the value is 0.5 meaning that, if the case gives the value of prob(conflict) in equation (3.2) greater than 0.5 it is classified in the conflict category and if the value is less than 0.5 we see the case as non conflict. However, for class imbalance problems the threshold may be increased or decreased accordingly.

Generally,

$$\text{Decision}=\begin{cases} Conflict \quad if\ prob(conflict) > 0.5 \\ No\ Conflict\ if\ prob(conflict) \leq 0.5 \end{cases}$$

Stepwise logistic regression was carried out to find best set of significant variables and final model was then built. The performance of the classifier was tested using ROC curve and confusion matrix.

### 3.4.3 Ordinal logit model for severity of P-V conflict.

Ordinal logit model is similar to binary logit, the only difference being that the dependent variable which was dichotomous (0 and 1), is now ordered consisting of more than than two levels. It is a statistical analysis method that can be used to model the relationship between an ordinal response variable and one or more explanatory variables. An ordinal variable is a categorical variable for which there is a clear ordering of the category levels. The explanatory variables may be either continuous or categorical. A major assumption of ordinal logistic regression is the assumption of proportional odds: the effect of an independent variable is constant for each increase in the level of the response. Hence the output of an ordinal logistic regression will contain an intercept for each level of the response except one, and a single slope for each explanatory variable.

Level of Severity of conflict has been defined as a variable related to vehicle speed and safety margin. Safety margin alone cannot alone define the severity of conflict as the severity of conflict resulting into a potential crash also depends upon vehicle speed. Therefore, a new indicator of safety, Risk indicator(RI) which is the ratio of vehicle speed to safety margin has been used quantify the severity of the potential crash. Cumulative

distribution function(CDF) of RI at 25%, 50%, 75% and 100% respectively was used to derive four levels of severity.

Mathematically,

if a latent variable, $Y^* = \beta_0 + \beta_1 X_1 + \beta_2 * X_2 + \ldots + \beta_k * X_k + \in$ ….(3.4)

In this particular case, for four levels, two cutoff points ($\alpha_1, \alpha_{2,}, \alpha_{3,}$) will be determined such that,

$$\text{Decision} = \begin{cases} No\ Risk\ if\ Y^* < \alpha_1 \\ Slight\ Risk\ if\ \alpha_2 > Y^* > \alpha_1 \\ Fair\ Risk\ if \alpha_3 > Y^* > \alpha_2 \\ High\ Risk \text{ if } Y^* > \alpha_3 \end{cases}$$

**Assumptions**

Some common assumptions before developing the model are :-

1)The independent variables must be least related to each other.

2)The explanatory variables should have linear relationship with logit of response variable.

3)The response variable is binary (for binary logit) and ordered (for ordinal)

4) Observations are independent.

5)The residuals (error) of the model shall be normally distributed.

**Parameter Estimation of the models.**

For binary logit model, there are iterative procedures and heuristic methods to estimate the parameters by maximizing the log likelihood, however, ***glm() and clm()*** functions in R directly gives us the parameters and errors associated ,by using their in built algorithms. After developing preliminary models stepwise regression was performed to find best set of significant variables for final model.

### 3.4.4 Model Validation

After creating a model , the performance of it should be tested using entirely different set of data to confirm the model is as good for broad data set. Model validation increases the scalability and flexibility of model, reduces underfit and overfit, discovers more errors to enhance the overall quality of model. An **underfit model** results in high prediction errors for both training and test data. An **overfit model** gives a very low prediction error on training data, but a very high prediction error on test data. Both types of models result in poor accuracy.

The dataset was first divided into training set and testing set randomly. Training set was used for building model while testing set was used to validate it. Goodness of fit or performance of model was determined using following techniques:

**1) R-squared value**

R-squared evaluates the scatter of the data points around the fitted regression line. It is also called the coefficient of determination, or the coefficient of multiple determination for multiple regression. For the same data set, higher R-squared values represent smaller differences between the observed data and the fitted values.
R-squared is the percentage of the dependent variable variation that a linear model explains.

$$R^2 = \frac{Variance\ explained\ by\ model}{Total\ variance}$$

When the value is 1, the model becomes a perfect fit.

**2) Confusion Matrix**

For measuring performance of a classifier, especially for categorical/ordinal response variable , confusion matrix is a common technique. A confusion matrix is a table that will categorize the predictions against the actual values. It includes two dimensions; among them one will indicate the predicted values and another one will represent the actual values.

If our model correctly predicts conflict for conflict and no conflict for no-conflict , there is no error. If the model predicts conflict for no conflict, there is a type I error (false positive); likewise if the model predicts no conflict for conflict, there exists type II error (false negative).The error cases are shown diagrammatically in Table 3.1.

**Table 3.1:-General Confusion Matrix**

| | | **Actual** | |
|---|---|---|---|
| | | Conflict | No conflict |
| **Predicted** | Conflict | True Positive | False Positive(Type I) |
| | No conflict | False Negative(Type II) | True Negative |

Various indicators derived from the matrix are:

**1) Accuracy**

Accuracy (ACC) is calculated as the number of all correct predictions divided by the total number of the dataset. The best accuracy is 1.0, whereas the worst is 0.0.It is sum of values in the diagonals divided by sum of values of all cells in the matrix.

**2) Sensitivity (Recall or True positive rate)**

Sensitivity (SN) is calculated as the number of correct positive predictions divided by the total number of positives. It is also called recall (REC) or true positive rate (TPR). The best sensitivity is 1.0, whereas the worst is 0.0.

**3) Specificity (True negative rate)**

Specificity (SP) is calculated as the number of correct negative predictions divided by the total number of negatives. It is also called true negative rate (TNR). The best specificity is 1.0, whereas the worst is 0.0.

**4) Precision (Positive predictive value)**

Precision (PREC) is calculated as the number of correct positive predictions divided by the total number of positive predictions. It is also called positive predictive value (PPV). The best precision is 1.0, whereas the worst is 0.0.

**3) ROC curve**

It is the plot of True Positive Rate and False Positive Rate for different thresholds. As threshold increases the true positive rate decreases; if the false positive rate also decreases but at a lower rate then the classifier has poor performance. But if the true positive rate increases with less increase in false positive rate as threshold increases, the classifier is said to be performing good. For different models, AUC (Area Under ROC Curve) gives the performance measure: AUC about 0.7-0.8 makes the model acceptable, 0.8 to 0.9 is excellent and greater than 0.9 is outstanding.

### 3.5 Sensitivity Analysis

After a model is built, sensitivity analysis is performed to examine how the response (here, probability of conflict/levels of severity of conflict) for a particular predictor varies across another predictor with other independent variables of the model remaining fixed. This type of analysis plot also known as Partial Dependence Plot(PDP), helps to depict scenarios of how two or more independent variables are related to each other with respect to the response (a method of sensitivity analysis). Such plots provide a clear picture and are often superior to marginal effects analysis for complex models as in these types of models the marginal effect varies across the values of a predictor; and average of these marginal effect values(AME) doesnot speak for these variations. PDPs for various scenarios was obtained and analyzed, for relatively significant predictors of model.

### 3.6 Sample Size Determination

There are many ways to determine the sample size, in this particular case the population size is unknown so rule of thumb has been used to calculate the sample size. Green (1991) recommended minimum sample size of N ≥ 50+8*m (where m is the number of predictors).According to this formula,

Sample Size >= 50 +8*15 >170

15 is the estimated no of independent variables.

Another commonly used method is :

$$Sample\ Size >= \frac{z^2 p(1-p)}{E^2}$$

where,

- n is the required sample size.
- Z is the z-score corresponding to the desired confidence level. For example, for a 95% confidence level, the z-score is approximately 1.96.
- p is the estimated proportion of the population that possesses the characteristic of interest. If you don't have an estimate, you can use 0.5 for a conservative estimate.
- E is the desired margin of error.

Assuming a desired confidence level of 95% (z-score = 1.96), an estimated proportion of 0.5, and a margin of error of 5%. Plugging these values into the formula, we can calculate the sample size:

n = $\frac{(1.96)^2 * 0.5 * (1-0.5)}{0.05\text{^}2}$ = 384.16

For this study, more than 1200 sets of data were collected from four crossings.

### 3.7 Data Collection

After identification of sites, videographic survey was conducted using a Go-Pro camera. The camera was placed at suitable height to capture the range of all pedestrian vehicle interaction throughout the crossings and, reference marks such that the vehicle speed

could be calculated from the video. Later, the recording was played on VLC media player for frame by frame extraction of data. To ensure, randomness in observation, the extraction, in general, was done for every 3rd pedestrian. survey was done during day time 9:30-11:00 am hours and 4:00pm to 5:00pm hours on weekdays under ideal weather condition. The time for survey was chosen from pilot study in such a way that high volume of pedestrian-vehicle interactions could be obtained. Certain observations were excluded from the study:-

1) The interaction on north and south bound lanes were excluded because of side friction .
2) Any signal by pedestrians to yield vehicles, affects the free interaction, so these datasets were excluded.
3)Data for pedestrians crossing at some angle with the road(not perpendicular) was ignored.

### 3.8 Data Extraction

The video was replayed and was analyzed on frame by frame basis on VLC media Player. For measuring speed of vehicles, time taken by them to cross certain reference marks was noted down to be later divided by the distance between the marks. Observation started as soon as the pedestrian finished crossing the first part and entered the shelter before crossing the middle lane. Time as he/she arrives at the shelter, departs from the shelter, reaches the end of lanes, departs from the lane was noted. Pedestrian age, gender, group-size, vehicle type and yielding of vehicles was visually observed from the video. Pedestrian speed, waiting time, vehicle gap was calculated based on the times noted on observation sheet. The data was obtained for every third pedestrian crossing the road to ensure randomness and avoid bias.

Measurements were first noted down on a primary observation sheet which was refined in many stages to obtain a clean dataframe. This cleansing was done in excel and later in R which involved omission of NA values, separation of categories, identification and editing of miscategorized datasets and removal of outliers. Outliers for various variables were removed by inter-quartile method.

The summary of data with median and IQR values obtained is given below in Table 3.2:-

**Table 3.2:-Data Information**

| Variables | **N = 1229** | Variable Type |
|---|---|---|
| Age | | |
| <20 | 289 (23%) | Categorical |
| 20-40 | 645 (52%) | |
| >40 | 294 (25%) | |
| GR_Size | | |
| 1 | 363 (29%) | Categorical |
| 2 | 332 (27%) | |
| (3-4) | 306 (24%) | |
| (5-7) | 228 (18%) | |
| Gen | | |
| F | 475 (38%) | Categorical |
| M | 754 (61%) | |
| Dry | | |
| 0 | 744 (60%) | Categorical |
| 1 | 250 (20%) | |
| 2 | 234 (19%) | |
| Wait_Time (s) | 4 (1, 11) | Continuous |
| Ped_Speed (m/s) | 1.09 (0.92, 1.33) | Continuous |
| PTC (s) | 3.00 (2.50, 3.51) | Continuous |
| Veh_gap (s) | 5.8 (4.0, 8.8) | Continuous |
| SM (s) | 1.8 (0.6, 3.9) | Continuous |
| Veh_typ | | |
| Two Wheeler | 872 (71%) | Categorical |
| Car_Jeep_VAN | 294 (24%) | |
| Heavy | 61 (5.1%) | |
| Veh_Speed (m/s) | 10.6 (8.0, 13.5) | Continuous |
| lane | | |
| 1 | 307 (25%) | Categorical |
| 2 | 307 (25%) | |
| 3 | 307 (25%) | |
| 4 | 308 (25%) | |
| LE_ILLE | | |
| Illegal | 639 (52%) | Categorical |
| Legal | 589 (48%) | |

### 3.9 Data Cleansing

Raw data obtained through video extraction was cleansed at multiple levels to facilitate model development. N/A values and other inconsistent dataset was manually identified and removed from Excel sheet. The outliers were then removed using interquartile range (IQR) method in R.

$$\text{IQR} = 3^{\text{rd}}\ \text{Quartile}(\text{Q}_3) - 1^{\text{st}}\text{Quartile}(\text{Q}_1) \quad \ldots(3.5)$$

$$Lower\ bound = Q_1 - 1.5 * IQR \quad \ldots(3.6)$$

$$Upper\ bound = Q_3 + 1.5 * IQR \ldots(3.7)$$

Q1 is a value for a variable for which 25% of the variable's data values are less than or equal to that particular value. Likewise, 75% of the variable's data values is less than Q3. The lower bound and upper bound , as defined in equations above, of continuous variables namely, waiting time, vehicle gap, safety margin, pedestrian speed and vehicle speed were evaluated and their corresponding dataset were deemed outliers. Remaining N/A values were also eliminated simultaneously.

Furthermore, categorical variables like group size, gender, yielding behaviour of driver, vehicle type, lane type, type of crossing and conflict were divided into suitable classes, labeled, and then converted to factor for further data analysis in R. Other continuous variables like wait time, pedestrian speed, safety margin, vehicle gap and vehicle speed were treated as numeric data type.

# CHAPTER 4 : RESULTS AND DISCUSSION

## 4.1 Correlation Matrix

The dataset collected for this study contains both continuous and categorical datatypes. From literature review, it was found that one particular method couldnot be used for correlation analysis in case of mixed dataset. Pearson Chi-Squared is suggested for correlation analysis of categorical-categorical variables, similarly, point-biserial test for dichotomous categorical and continuous variables and Pearson correlation test for continuous-continuous dataset are recommended. For our dataset, Pearson correlation and Cramers V tests were done for numeric datatypes and categorical datatypes respectively in R whose results have been summarized below.

**Pearson Correlation Test Results for numeric datatypes:**

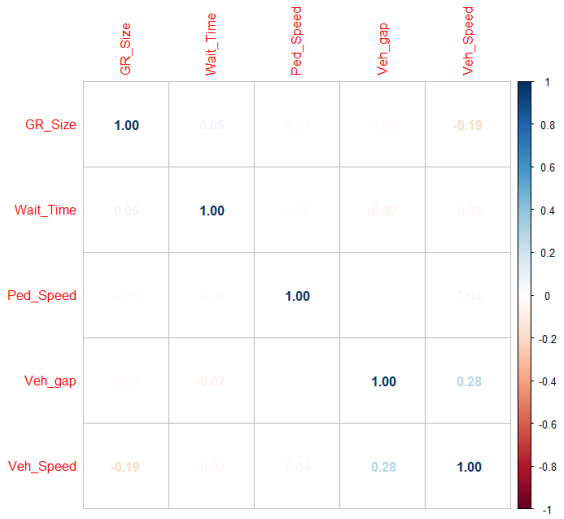


**Figure 4.1: Correlation map for numeric datatypes**

**Table 4.1: Correlation Matrix for numeric independent variables**

| | GR_Size | Wait_Time | Ped_Speed | Veh_gap | Veh_Speed |
|---|---|---|---|---|---|
| GR_Size | 1.00 | 0.05 | -0.01 | -0.02 | -0.19 |
| Wait_Time | 0.05 | 1.00 | 0.02 | -0.07 | -0.03 |
| Ped_Speed | -0.01 | 0.02 | 1.00 | 0.00 | 0.04 |
| Veh_gap | -0.02 | -0.07 | 0.00 | 1.00 | 0.28 |
| Veh_Speed | -0.19 | -0.03 | 0.04 | 0.28 | 1.00 |

From the results depicted in Figure 4.1 and Table 4.1, we can see that there are no strong correlations between two numeric independent variables. Literatures suggest correlation coefficients greater than 0.5 as strong (multicollinear) and the need to eliminate one of the two variables before model development in such case. Muticollinearity decreases significance of variables and overall model. Duplication of predictors also makes the model vague, imprecise and absurd.

**Cramers V Correlation Test Results for categorical datatypes:**

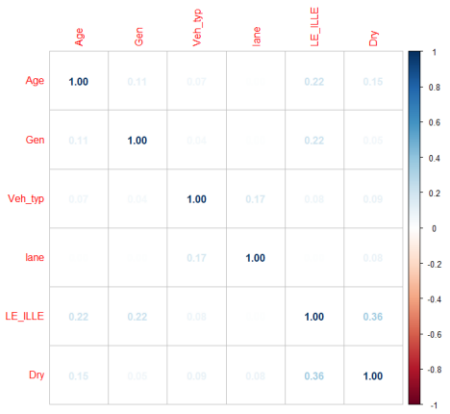


**Figure 4.2: Correlation map for categorical datatypes**

**Table 4.2: Cramers V Correlation Matrix for categorical independent variables**

| | Age | Gen | Veh_typ | lane | LE_ILLE | Dry |
|---|---|---|---|---|---|---|
| Age | 1.00 | 0.11 | 0.07 | 0.00 | 0.22 | 0.15 |
| Gen | 0.11 | 1.00 | 0.04 | 0.00 | 0.22 | 0.05 |
| Veh_typ | 0.07 | 0.04 | 1.00 | 0.17 | 0.08 | 0.09 |
| lane | 0.00 | 0.00 | 0.17 | 1.00 | 0.00 | 0.08 |
| LE_ILLE | 0.22 | 0.22 | 0.08 | 0.00 | 1.00 | 0.36 |
| Dry | 0.15 | 0.05 | 0.09 | 0.08 | 0.36 | 1.00 |

Similarly, the categorical variables, like numeric, show no strong correlation with each other.

## 4.2 Results on Pedestrian -Vehicle Conflict (PVC) Model

Preliminary dataset from excel was imported and converted into dataframe via ***read.xlsx()*** function in R 4.3.3 software. The dataframe was cleaned in many stages; in first stage the data was merged into suitable format after which in second stage outliers were removed using interquartile method.In the third stage, N/A values were omitted from the dataframe. Categorical variables were converted into factors with suitable levels. A new column "conflict" was added to the cleaned dataframe whose value was dependent upon the SM values of each dataset; SM more than 1 would make conflict 0 whereas SM less than one would make conflict 1.

The final dataset was split into training (75%) and testing set (25%) randomly using ***catools()*** library.Training dataset was then used to build a preliminary model considering all variables which is shown in table 4.3 below.

**Table 4.3: PV Conflict Preliminary Model I summary:**

| | Dependent variable: |
|---|---|
| | log(odds_conflict) |
| Veh_Speed | -0.04 |
| | (0.05) |
| lane2 | 2.14*** |
| | (0.46) |
| lane3 | 1.07** |
| | (0.46) |
| lane4 | 2.34*** |
| | (0.44) |
| Veh_gap | -0.87*** |
| | (0.10) |
| Ped_Speed | -1.89*** |
| | (0.50) |
| Wait_Time | -0.04 |
| | (0.03) |
| GenM | -0.33 |
| | (0.31) |
| Veh_typCar_Jeep_VAN | -0.55 |
| | (0.35) |
| Veh_typHeavy | 0.48 |
| | (0.70) |
| GR_Size2 | 0.82** |
| | (0.37) |
| GR_Size(3-4) | 0.62 |
| | (0.42) |
| GR_Size(5-7) | -0.06 |
| | (0.54) |
| LE_ILLEIllegal | 0.06 |
| | (0.34) |
| Dry1 | 0.90* |
| | (0.46) |
| Dry2 | -0.14 |
| | (0.55) |
| Age20-40 | -1.05*** |
| | (0.39) |
| Age> 40 | -0.13 |
| | (0.50) |
| Constant | 5.72*** |
| | (1.06) |
| Observations | 829 |
| Log Likelihood | -163.32 |
| Akaike Inf. Crit. | 364.64 |
| Residual Deviance | 326.64 (df = 810) |
| Null Deviance | 614.76 (df = 828) |
| Note: | *p<0.1; **p<0.05; ***p<0.01 |

PV Conflict Model
====
TRUE

Model I summary shows variables namely, vehicle speed, wait time, gender, vehicle type, legal/illegal to be insignificant at 95% confidence level. Significance of variables exists at certain level when the null hypothesis stating that the relationship between the predictor and dependent variables exists by chance is proven false at that very level of significance. In other words, the model with coefficient of the significant variable adds more value than without it. So, stepwise regression is carried out in both forward and backward direction to find out the best set of significant variables. The results of stepwise regression models also constituted of many insignificant variables. After multiple trials, two sets of models having lowest AIC values and maximum log likelihood values and also with inclusion of important variables were obtained and presented below in table 4.4.

- Model (1) is the final result of the stepwise regression (both direction).
- Model (2) and (3) are the results of multiple trials with an attempt to include only significant variables. While Model (3) includes age and yielding behavior of vehicles and excludes waiting time, Model (2), on the other hand, excludes both age and yielding behavior and includes waiting time. Additionally, compared to Model (3), Model (2) has more significant variables but lower log-likelihood and AIC values.

The coefficients of most variables in all PVC models are approximately similar to each other. From the results it can be seen that, pedestrians have higher chances of conflict while crossing second lane and fourth lane compared to first lane. The probability of odds of such a conflict increases nearly by 1.97 times in second lane and even more, by 2.29 times in fourth lane (last lane), while for the third lane the odds increases but by a lesser amount, nearly 0.95 times. Availability of shelters at the beginning and ending of crossings provided comfortable waiting zones for pedestrians. Also, while pedestrians did not wait that much at the middle of the road; they did it much frequently and much longer than at the ending of the first and third lanes. Thus, it can be seen that if the pedestrians get comfortable waiting shelters, they are less inclined to show conflicting behaviour or the probability of conflict decreases. This also shows the risk pedestrians faces at wide crossings of multiple lanes with no proper resting place.

Acceptance of a unit large vehicular gaps is shown to decrease the odds of probability of conflict by 0.8 times. Similarly, pedestrian speed also has negative correlation with conflict with a coefficient of -1.93; increase in one unit of speed decreases the probability of conflict nearly by two times. Furthermore, pedestrians who spend more time at the road and the shelter (waiting time) are found to have less conflicting chances. Waiting time is negatively correlated with conflict by 0.05.

As pedestrian cross in groups, they tend to show more conflicting behavior than crossing alone. When the group size is 2 the odds of showing conflicting behavior increases by 0.81 time compared to single pedestrians. Similarly, when the group size if 3-4 such odds is found to increase but by a lesser amount 0.78. However, when crossing in larger groups of 5-7, pedestrians show nearly same (0.04 in model 2) or even less conflicting behaviour (-0.27 in model 3).

Model 3 shows pedestrians of age group 20-40 were found to show less conflicting behavior than younger pedestrians of below 20 .The odds decreased by 1.11 for the 20-40 group while it decreased by a lesser amount 0.27 for pedestrians in greater than 40 age group, showing that the younger and older pedestrians show less risking behavior than pedestrians of middle age groups.

**Table 4.4 : PV Conflict Final Models Summary**

| | Dependent variable: | | |
|---|---|---|---|
| | log(odds_conflict) | | |
| | (Model 1) | (Model 2) | (Model 3) |
| Veh_Speed | -0.01 | | |
| | (0.05) | | |
| lane2 | 2.01*** | 1.97*** | 1.93*** |
| | (0.44) | (0.41) | (0.41) |
| lane3 | 1.18*** | 1.09*** | 0.95** |
| | (0.44) | (0.41) | (0.42) |
| lane4 | 2.49*** | 2.29*** | 2.29*** |
| | (0.43) | (0.40) | (0.41) |
| Veh_gap | -0.85*** | -0.80*** | -0.85*** |
| | (0.09) | (0.08) | (0.09) |
| Ped_Speed | -1.81*** | -1.93*** | -1.93*** |
| | (0.48) | (0.44) | (0.46) |
| Wait_Time | -0.04 | -0.05** | |
| | (0.03) | (0.02) | |
| GenM | -0.31 | | |
| | (0.29) | | |
| Veh_typCar_Jeep_VAN | -0.55 | | |
| | (0.35) | | |
| Veh_typHeavy | 0.65 | | |
| | (0.59) | | |
| GR_Size2 | 0.87** | 0.81** | 0.84** |
| | (0.36) | (0.34) | (0.35) |
| GR_Size(3-4) | 0.73* | 0.78** | 0.68* |
| | (0.40) | (0.37) | (0.38) |
| GR_Size(5-7) | -0.13 | 0.04 | -0.27 |
| | (0.52) | (0.41) | (0.49) |
| LE_ILLEIllegal | -0.13 | | |
| | (0.32) | | |
| Dry1 | 1.06** | | 0.87** |
| | (0.45) | | (0.40) |
| Dry2 | -0.03 | | 0.02 |
| | (0.52) | | (0.41) |
| Age20-40 | -0.91** | | -1.11*** |
| | (0.37) | | (0.35) |
| Age> 40 | -0.03 | | -0.27 |
| | (0.47) | | (0.45) |
| Constant | 5.16*** | 4.28*** | 4.99*** |
| | (0.99) | (0.68) | (0.82) |
| Observations | 829 | 829 | 829 |
| Log Likelihood | -177.35 | -188.74 | -181.75 |
| Akaike Inf. Crit. | 392.69 | 397.47 | 389.51 |
| Residual Deviance | 354.69 (df = 810) | 377.47 (df = 810) | 363.51 (df = 810) |
| Null Deviance (df = 828) | 658.03 | 658.03 | 658.03 |
| Note: | | | *p<0.1; **p<0.05; ***p<0.01 |

PV Conflict Model
====
TRUE

### 4.3 Kruskal-Wallis Test on Severity Levels

To test whether the levels defined are different from each other or not, Kruskal Wallis test was performed against 5 independent numeric variables one by one. Kruskal Wallis test is a standard nonparametric method for testing whether samples are originated from the same distribution. While ANOVA test is used for categorical independent variables, this test is used when the dependent variable is categorical. The hypothesis of the test is stated as below:

***Null Hypothesis***: The medians (mean on ranks) are equal across the levels.
***Alternative Hypothesis***: At least one median is different.

The results are tabulated in Table 4.5 below:

**Table 4.5 : Kruskall-Wallis Test Results**

| S.N | Variables | Chi-Squared | P value |
|---|---|---|---|
| 1 | Vehicle Gap | 629.12 | 0.00 |
| 2 | Vehicle Speed | 467.75 | 0.00 |
| 3 | Safety Margin | 612.5 | 0.00 |
| 4 | Waiting Time | 297.13 | 0.00 |
| 5 | Pedestrian Speed | 485.44 | 0.00 |

Results show the four levels defined donot have same distribution and are essentially different from each other as p-value for chi squared test is nearly zero allowing us to reject the null hypothesis.

### 4.4 Results on Pedestrian -Vehicle Conflict Severity Model

A new column of PVSRI was added in the dataset as the ratio of vehicle speed to safety margin. To indicate that increasing PVSRI corresponds with increasing severity, the lowest range of safety margin, which was previously negative, was now shifted to zero. Dataset was again filtered to remove outliers based on new PVSRI variable. Four levels of severity were also defined and added. This modified dataset was split into training

(75%) and testing set (25%) as mentioned before. The former was used to create a model while the latter was used for validation through confusion matrix. Considering all the variables we obtain our primary model as follows in Table 4.6:-

**Table 4.6 : Preliminary Severity Model I Summary**

Ordinal Regression Output

| | Dependent variable: |
|---|---|
| | PVSRI |
| lane2 | 1.52*** |
| | (0.27) |
| lane3 | 1.11*** |
| | (0.26) |
| lane4 | 1.90*** |
| | (0.27) |
| Veh_gap | -0.47*** |
| | (0.04) |
| Ped_Speed | -1.48*** |
| | (0.30) |
| Wait_Time | -0.03** |
| | (0.01) |
| GenM | -0.09 |
| | (0.20) |
| Veh_typCar_Jeep_VAN | -0.76*** |
| | (0.23) |
| Veh_typHeavy | 0.02 |
| | (0.38) |
| GR_Size2 | -0.17 |
| | (0.24) |
| GR_Size(3-4) | 0.03 |
| | (0.25) |
| GR_Size(5-7) | -0.09 |
| | (0.30) |
| LE_ILLEIllegal | 0.53** |
| | (0.21) |
| Dry1 | -0.33 |
| | (0.26) |
| Dry2 | -2.83*** |
| | (0.37) |
| Age20-40 | 0.63** |
| | (0.26) |
| Age> 40 | 0.55* |
| | (0.33) |
| Observations | 829 |
| Log Likelihood | -535.20 |
| Note: | *p<0.1; **p<0.05; ***p<0.01 |

Threshold coefficients:

| | Estimate | Std. Error | z value |
|---|---|---|---|
| No risk\|Slight risk | -4.9467 | 0.5742 | -8.615 |
| Slight risk\|Fair risk | -3.4168 | 0.5507 | -6.204 |
| Fair risk\|High Risk | -1.7852 | 0.5298 | -3.370 |

Variables namely, gender, group size, slow yielding behaviour of driver and heavy vehicle type were shown insignificant in the model. Stepwise regression yielded following Model (1) results which still showed slow yielding behaviour of driver and heavy vehicle type insignificant at 95% confidence level. After performing multiple trials, best model with all significant variables is shown as Model (3) in Table 4.7 below.

Unlike Model (1) and Model (3), model Model (2) includes age but excludes vehicle type. If we compare loglikelihood, Model (1) produced from stepwise regression has the least while containing some insignificant variables. Model (3) however, has all significant variables and medium loglikelihood value between Model (1) and Model (2) upon which following interpretation is based.

**Table 4.7: Severity Models Summary**

Ordinal Regression Output

| | Dependent variable: | | |
|---|---|---|---|
| | PVSRI | | |
| | (Model 1) | (Model 2) | (Model 3) |
| lane2 | 1.66*** | 1.30*** | 1.64*** |
| | (0.26) | (0.25) | (0.26) |
| lane3 | 0.96*** | 0.75*** | 1.01*** |
| | (0.26) | (0.25) | (0.26) |
| lane4 | 1.87*** | 1.35*** | 1.54*** |
| | (0.27) | (0.25) | (0.26) |
| Age20-40 | | 0.52** | |
| | | (0.22) | |
| Age> 40 | | 0.31 | |
| | | (0.28) | |
| Veh_gap | -0.50*** | -0.38*** | -0.41*** |
| | (0.04) | (0.03) | (0.04) |
| Ped_Speed | -1.08*** | -0.86*** | -0.69** |
| | (0.30) | (0.28) | (0.29) |
| Wait_Time | -0.02 | -0.03** | -0.03** |
| | (0.01) | (0.01) | (0.01) |
| Veh_typCar_Jeep_VAN | -0.70*** | | -1.07*** |
| | (0.22) | | (0.21) |
| Veh_typHeavy | -0.95** | | -1.38*** |
| | (0.46) | | (0.46) |
| LE_ILLEIllegal | 0.75*** | 1.00*** | 1.21*** |
| | (0.20) | (0.19) | (0.20) |
| Dry1 | -0.83*** | | |
| | (0.25) | | |
| Dry2 | -2.96*** | | |
| | (0.35) | | |
| Observations | 800 | 800 | 800 |
| Log Likelihood | -533.07 | -588.20 | -575.73 |

Note: *p<0.1; **p<0.05; ***p<0.01

Threshold coefficients of Model (1):

| | Estimate | Std. Error | z value |
|---|---|---|---|
| No risk\|Slight risk | -4.8044 | 0.5279 | -9.101 |
| Slight risk\|Fair risk | -3.2743 | 0.5003 | -6.545 |
| Fair risk\|High Risk | -1.6448 | 0.4802 | -3.425 |

Threshold coefficients of Model (2):

| | Estimate | Std. Error | z value |
|---|---|---|---|
| No risk\|Slight risk | -3.3879 | 0.4575 | -7.405 |
| Slight risk\|Fair risk | -2.1034 | 0.4412 | -4.767 |
| Fair risk\|High Risk | -0.6547 | 0.4282 | -1.529 |

Threshold coefficients of Model (3):

| | Estimate | Std. Error | z value |
|---|---|---|---|
| No risk\|Slight risk | -3.4959 | 0.4267 | -8.193 |
| Slight risk\|Fair risk | -2.1472 | 0.4099 | -5.238 |
| Fair risk\|High Risk | -0.6259 | 0.3966 | -1.578 |

Results show that the probability of higher severity of conflict is positively correlated with type of crossing. The odds of increase in severity increases by 1.21 (from Model (3)) in illegal crossings compared to legal ones with proper zebra and lane markings.

Models show that pedestrians take more severe risk while crossing second and final lanes. The result is consistent with the conflict model result in terms of the location of more conflicting lanes; odds of severe risk increases 1.64 times at the second lane, by 1.01 after pedestrian rests at the middle of the road and by 1.54 times at the final lane. This shows that pedestrians are willing to take more severe risk as number of lanes increases without availability of proper waiting shelters.

Availability of vehicle gap is negatively correlated with severity. A unit more gap is found to decrease the odds by 0.41 times. Similarly, increase in pedestrian speed and wait time are also found to decrease the odds by 0.69 and 0.03 times respectively. This shows pedestrians who wait more before crossing the road and spend more time on the road while crossing are found to have lesser risk of severe conflict. Also, the coefficients highlight that pedestrian who cross the road slowly are more safer.

Compared to younger age group of less than 20, pedestrians in age group 20-40 and greater than 40 are found to show more severe risking behavior. The odds increase by 0.75 times for 20-40 and by 0.68 for greater than 40 age groups.

As for vehicle type, two wheelers are found to be riskiest; the odds decreased by 1.07 and 1.38 times for cars_jeep_van and heavy vehicles like buses_trucks respectively. This may be because pedestrians perceived the threat of bigger vehicle size and kept more safety margin against heavy vehicles compared to the other two. They felt more safe and showed least risking behavior against cars, jeeps and vans.

### 4.5 Model Validation

The models presented in Table 4.6 were found to demonstrate similar nature in terms of values of coefficients and model properties (shown in Table 4.8); the only notable difference in them being in the significance of variables. Mc Fadden R squared value of 0.42 was delivered, which indicates model's good explanation for variation. It was validated using the test dataset which consisted of 25% of random original dataset.

**Table 4.8: General Properties PV Conflict Models**

| llh | llhNull | G2 | McFadden | r2ML | r2CU |
|---|---|---|---|---|---|
| -188. | -329.01 | 280.55 | 0.42 | 0.43 | 0.58 |

#### 4.5.1 Validation for PV Conflict Model

Model (2) of Table 4.6 was validated through confusion matrix and Receiver Operating Characteristic (ROC) curve.

**Table 4.9: Confusion Matrix of Model (2)**

| | | Actual | |
|---|---|---|---|
| | | **0** | **1** |
| Predicted | **0** | 96 | 17 |
| | **1** | 9 | 45 |

i) Accuracy = $\frac{(96+45)}{(96+45+17+9)}$ = 84.43%

ii) Sensitivity = 91.43%

iii) Specificity = 72.58%

iv) Positive Prediction value = 84.96%

The ROC curve at various thresholds for Model(2) is shown in figure below:

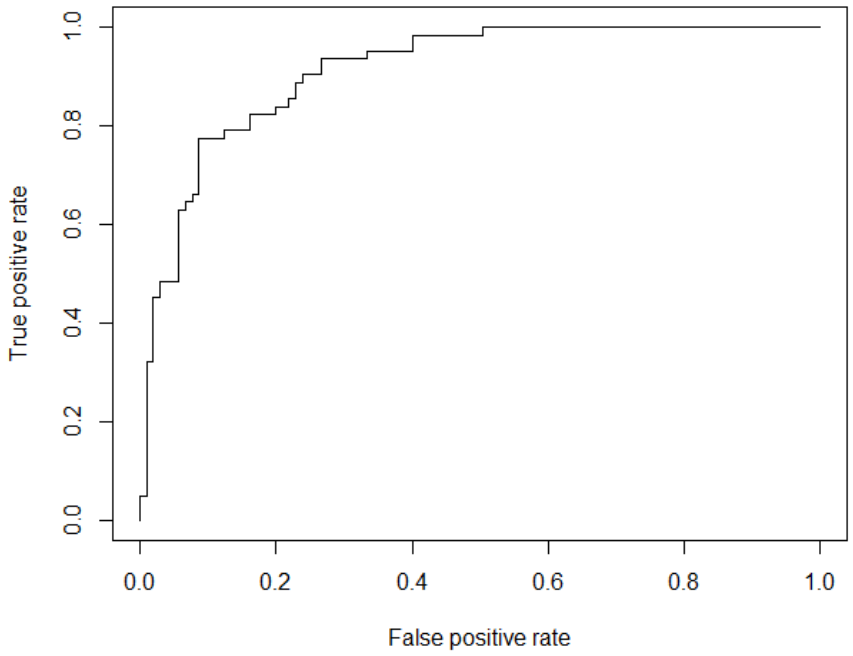


**Figure 4.3: ROC curve for PVC model (2)**

From the ROC curve, we can see that the true positive rate increases near vertical from 0.63 to 0.78 and 0.44 to 0.62 with no increment in the false positive rate values. The AUC value computed as the area under the ROC curve for the model is found to be 90.56 percent which shows excellent classification capability. Acquired R-squared value 0.43 is also considered good enough although it is not primary indicator of performance for logistic models. Validation results of other two models were similar to above and therefore, not discussed.

### 4.5.2 Validation for PV Severity Model

The test dataset was used for generating a confusion matrix to compute accuracy of all models; the best performance was shown by severity model (3) whose properties and confusion matrix are shown in Table 4.10 and 4.11 respectively.

**Table 4.10: Properties of Model (3)**

| link | threshold | nobs | logLik | AIC | niter | max.grad | cond.H | McFadden R2 |
|---|---|---|---|---|---|---|---|---|
| logit | flexible | 800 | -575.73 | 1175.45 | 5(0) | 1.57e-13 | 7 .9e+03 | 0.21 |

McFadden R2 squared value of 0.21 shows moderately good fit. McFadden's R2 values between 0.2 and 0.4 are taken to represent a very good fit of the model (McFadden, 1974). Simulations equivalence this range to 0.7 to 0.9 for a linear model (Louviere et al., 2000).

**Table 4.11: Confusion Matrix for Severity Model (3)**

| | | Actual | | | |
|---|---|---|---|---|---|
| | | **No Risk** | **Slight Risk** | **Fair Risk** | **Severe Risk** |
| Predicted | **No Risk** | 21 | 11 | 6 | 3 |
| | **Slight Risk** | 8 | 10 | 3 | 4 |
| | **Fair Risk** | 2 | 8 | 36 | 6 |
| | **Severe Risk** | 3 | 6 | 10 | 27 |

i) Accuracy = $\frac{(21+10+36+27)}{(21+10+36+27+8+10+3+15+6+14+11+6+3+3+4+11)}$ = 57.3%

The model could correctly classify 57.3 % of the test data to its respective severity level.

### 4.6 Sensitivity analysis through Partial Dependence Plots (PDP)

#### 4.6.1 PDPs for Pedestrian-Vehicle Conflict model

Probability of conflict was examined for one or two factors in a single plot while keeping other factors (marginal) at a constant mean (for continuous variables) or at a fixed category (for categorical variables).The values for factors of interest were extracted from the range of values it had in the training set. A dataset consisting of all possible combination of these values of factors of interest with the marginal factors was then created which was later fed into the model to receive the probability results. The results were plotted against the factors to visualize their interaction with respect to the response. Figure 4.4 shows the plot for pedestrian speed, vehicle gap and their predicted probability. It can be seen that for lower (near 0) and higher (near 15) vehicular gaps, with increase in pedestrian speed, there is no substantial decrease in the predicted probability; the decreasing trend is virtually negligible. However, for vehicular gap values between 5 and 10, the response is first decreasing in convex manner and transforming slowly to concave lines at higher vehicular gaps near 10, indicating that at that particular magnitude of vehicular gaps (mid of 5 and 10), the response decreases rapidly with increase in pedestrian speed.

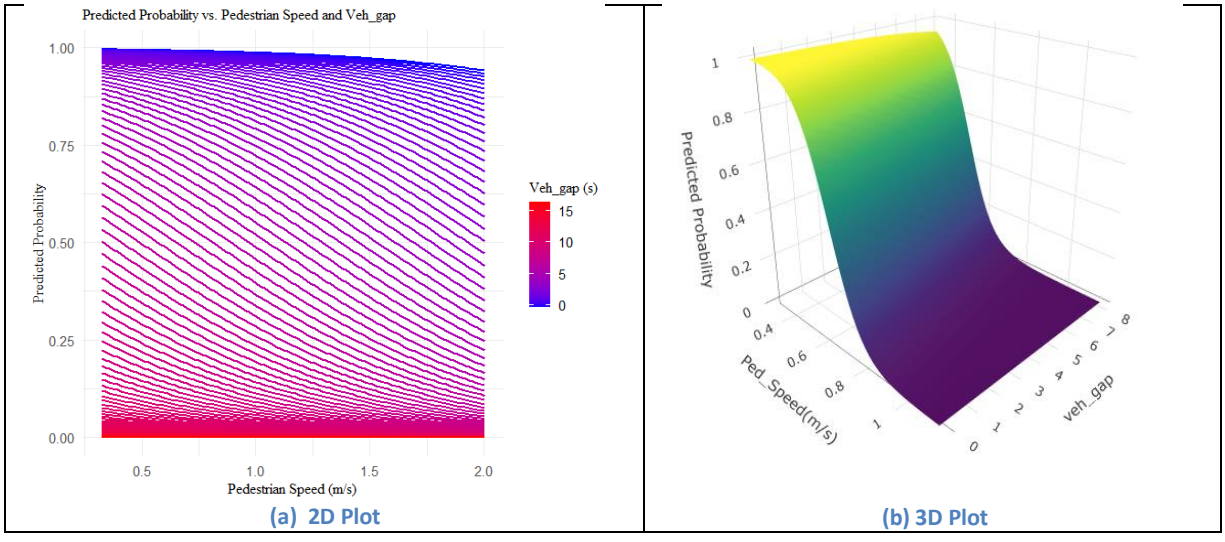


**Figure 4.4: PDP for Probability of conflict, Vehicular gap and Pedestrian Speed**

Likewise, it can observed from Figure 4.5 (a) which shows PDP for Pedestrian speed and Waiting time, that conflict probability decreases as speed increases and more importantly, the response is more sensitive to lower waiting times (less than 10s).The 3D plot in Figure 4.5(b) probability is highest for lower pedestrian speed and lower waiting times.

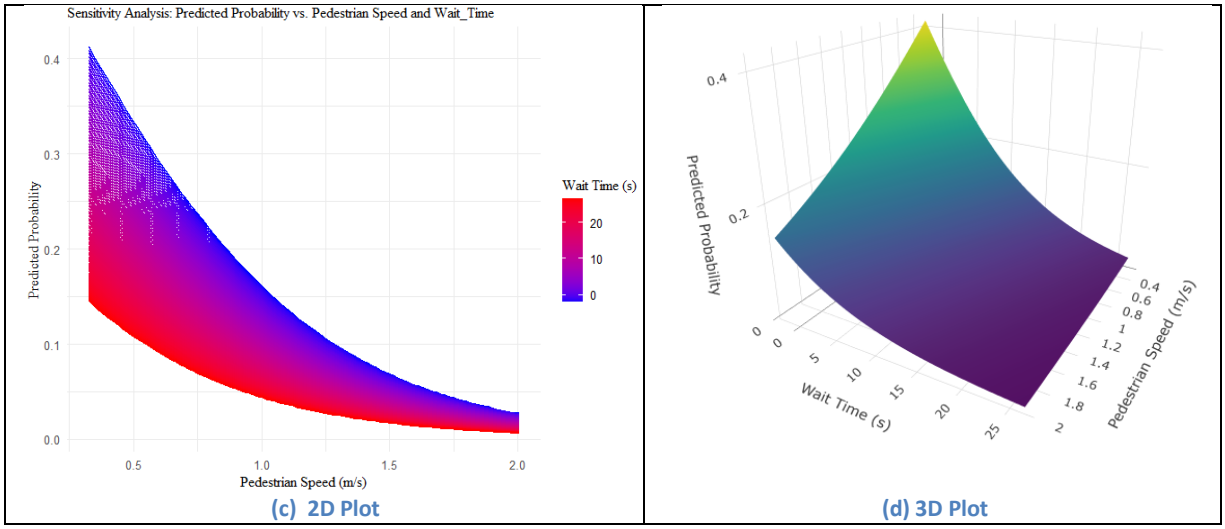


**Figure 4.5: PDP for Probability of conflict, Pedestrian Speed and Waiting Time**

Moreover, for similar vehicular gaps accepted by pedestrians, the probability of conflict is higher when they cross in group of 2 and 3-4, as depicted in Figure 4.6.The slope of the response plot changes sharply at two points: one near 5s vehicle gap and the other at near 8s gap while the lines tends to converge with increasing gap size. The plot indicates that pedestrians tend to take lower risk while crossing alone or in large groups of size 5-7.

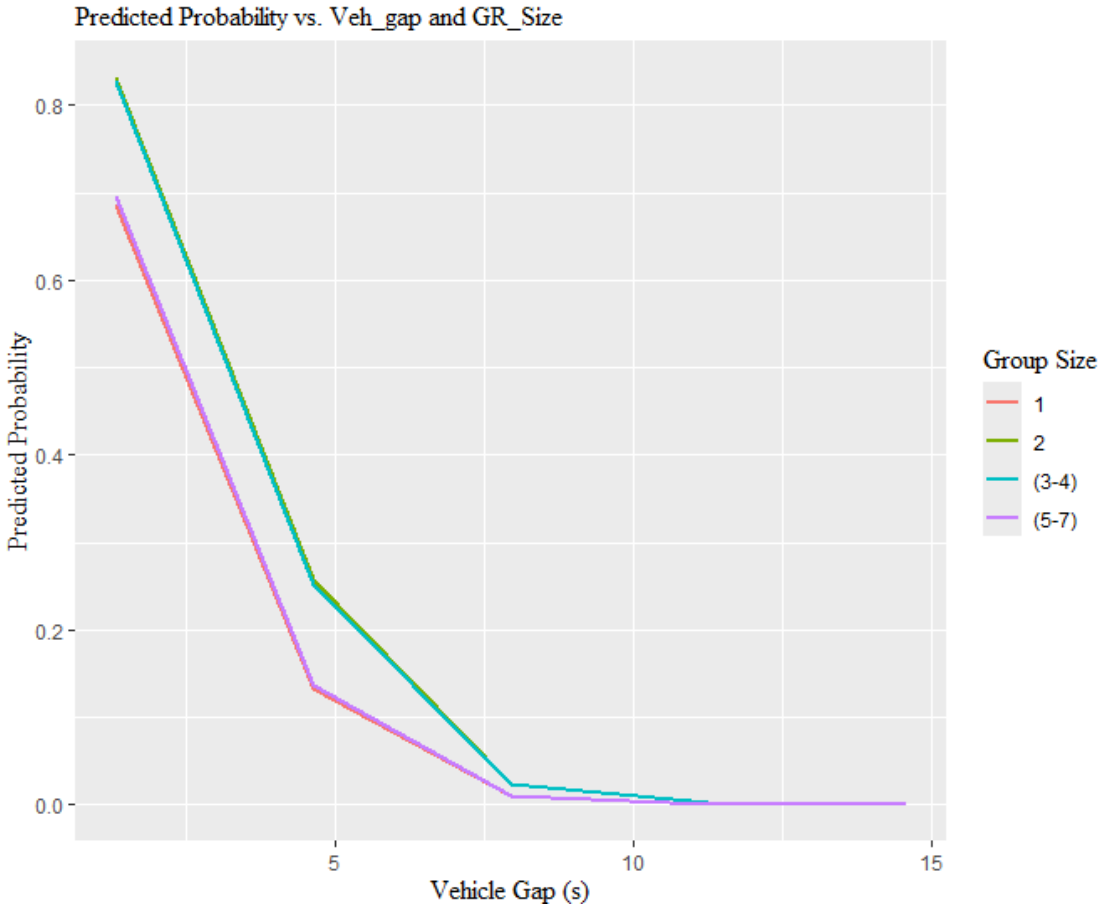


**Figure 4.6: PDP for Probability of conflict vs Accepted Vehicular Gap and Group Size.**

Likewise, from figure 4.7, the response for a given waiting time is highest for group size of 2 and a bit lower for 3-4 while being comparatively insensitive(shown by steady decrease of the plot line) for group size 1 and 5-7.

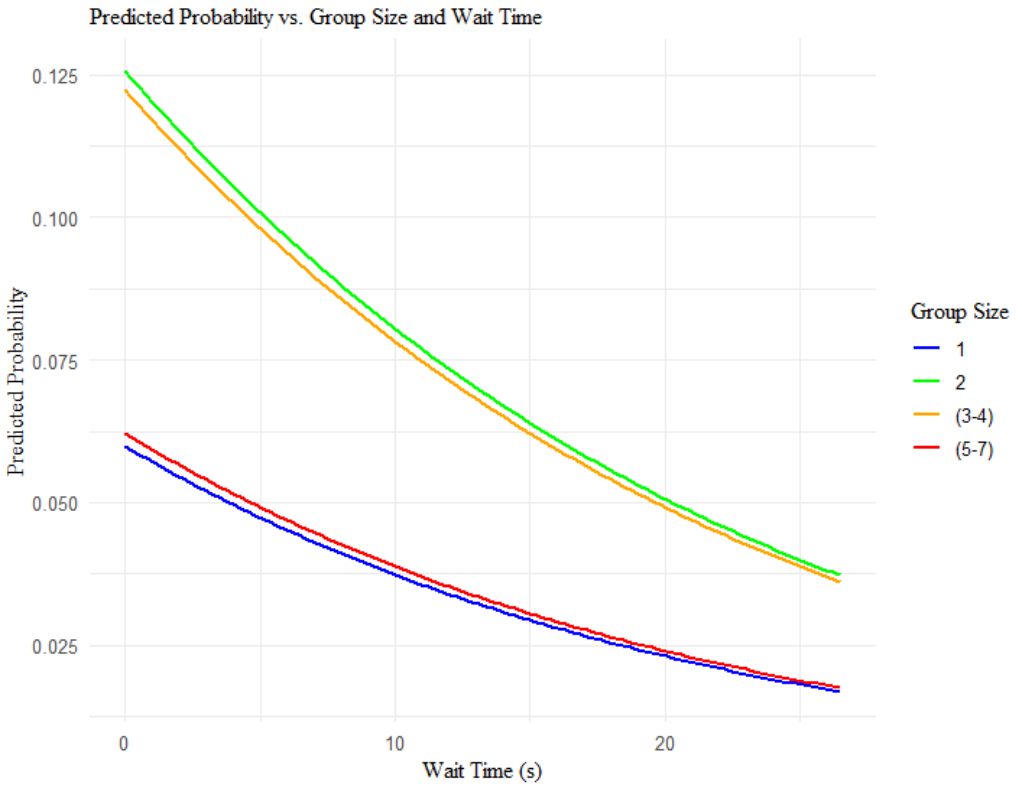


**Figure 4.7: PDP for Probability of conflict vs Wait Time and Group Size.**

Figures 4.8, 4.9, 4.10 reveal that probability of conflict is highest at final lane (lane 4) followed by lane 2, 3 and 1 respectively supporting the assertion provided by the model coefficients. Plot 4.8 and 4.9 are of similar nature in terms of suggesting that the sensitivity of the response is higher at lower values of pedestrian speed, vehicle gap and waiting time. In addition, the lines are highly sensitive to vehicular gap as seen in Figure 4.10.

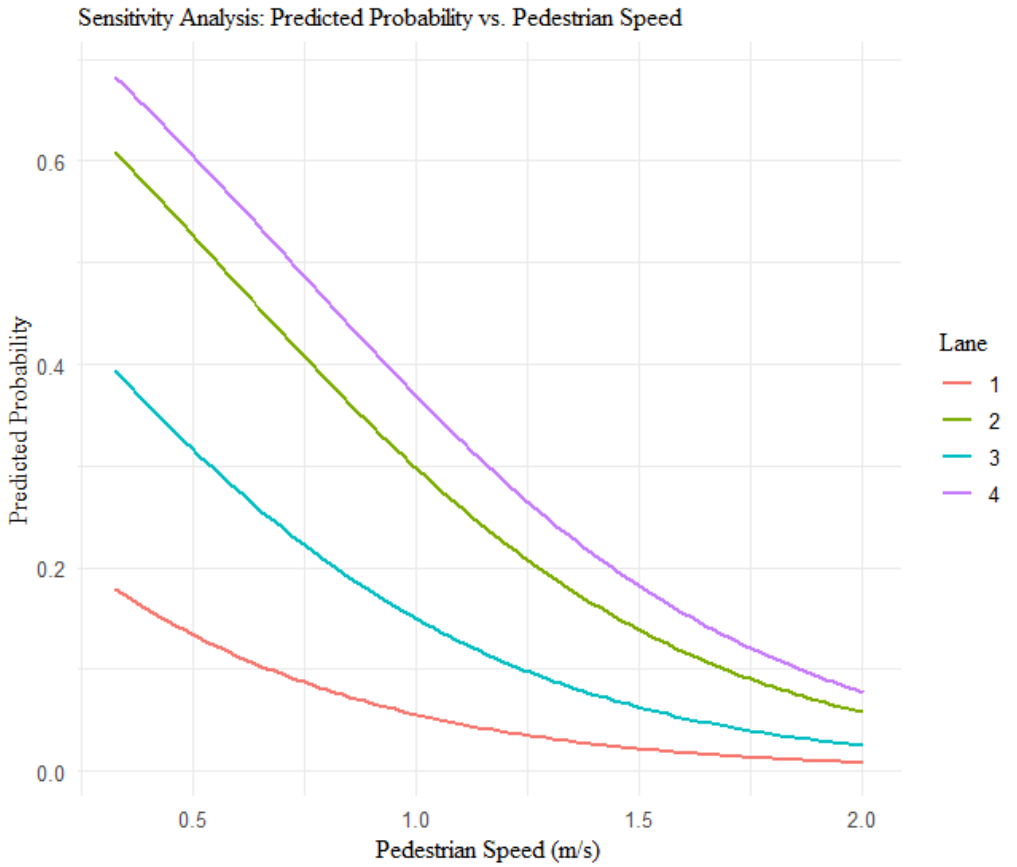


**Figure 4.8: PDP for Probability of conflict vs Lane and Pedestrian Speed.**

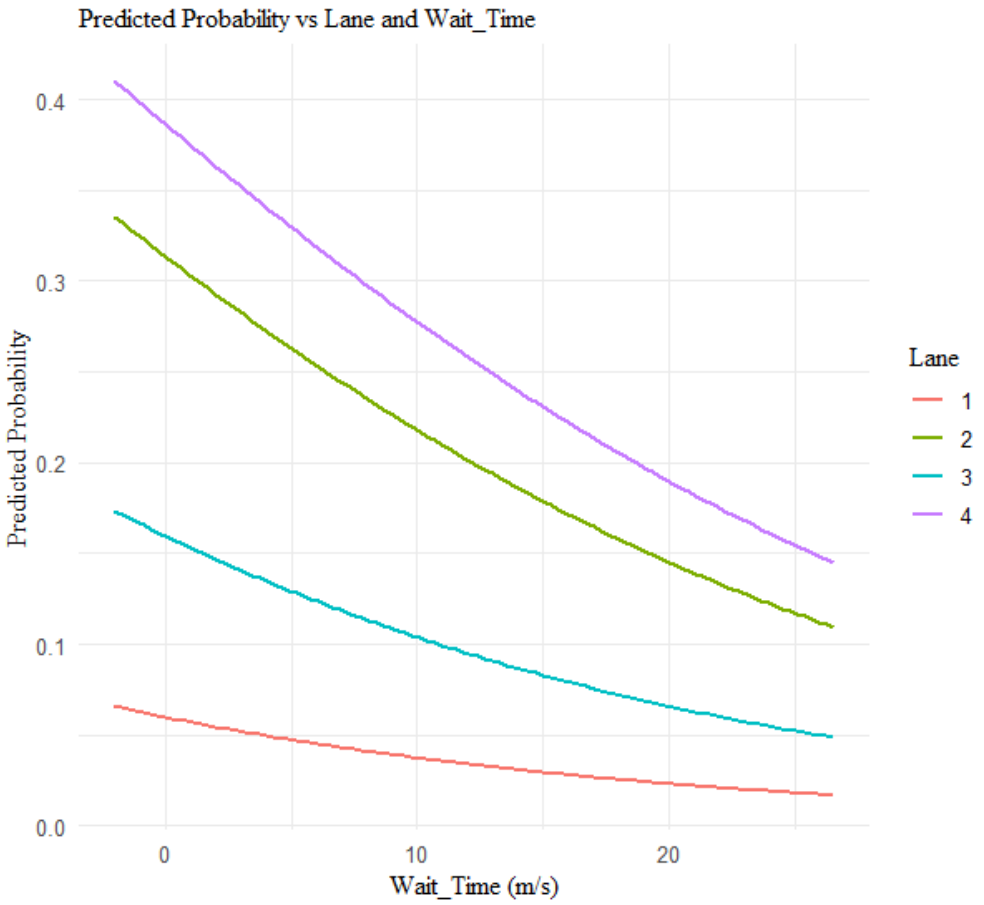


**Figure 4.9: PDP for Probability of conflict vs Lane and Wait Time.**

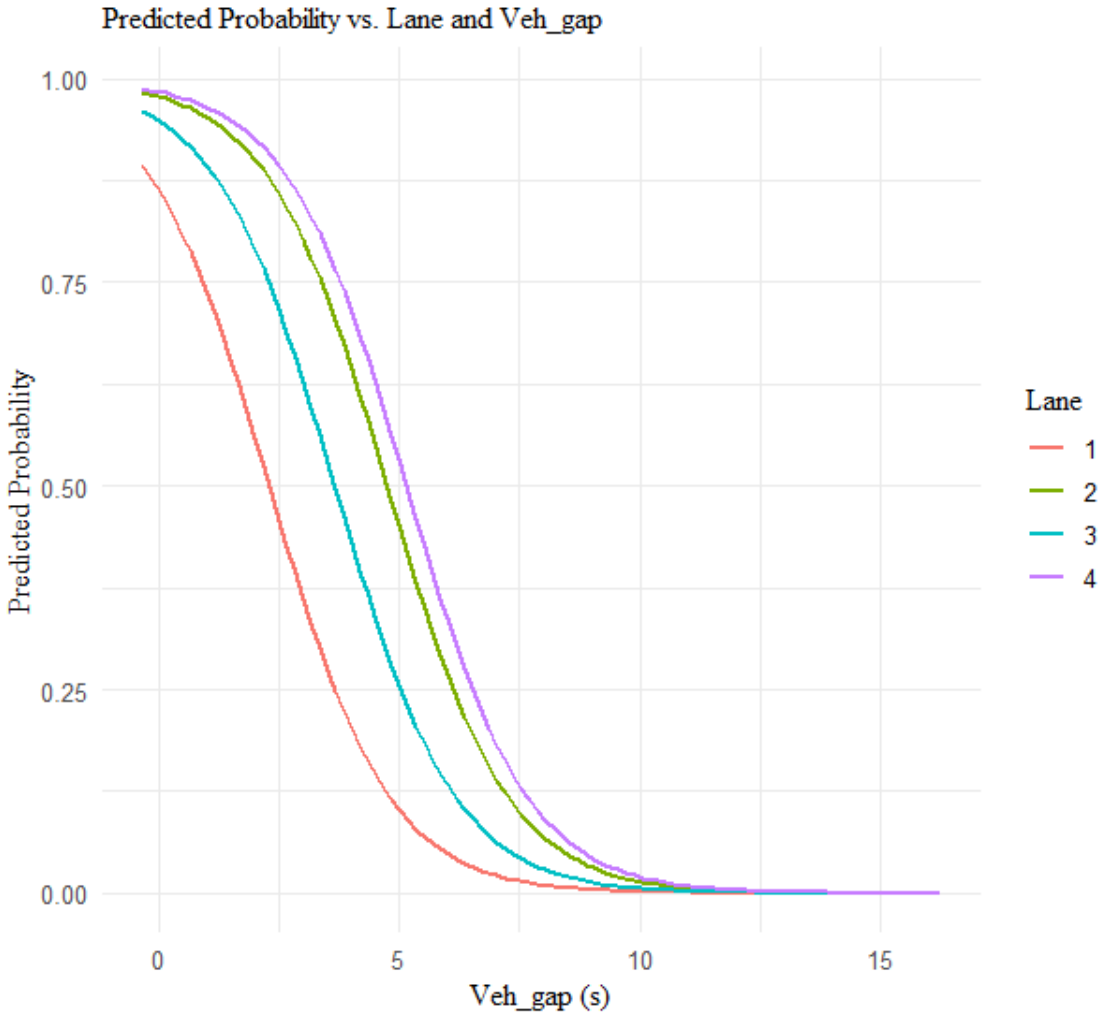


**Figure 4.10: PDP for Probability of conflict vs Lane and Vehicle gap.**

### 4.6.2 PDPs for Pedestrian-Vehicle Severity model

Plot for probability of certain level of severity of conflict at each values of wait time while keeping all other factors constant can be seen in Figure 4.11.The figure shows, validating the model coefficients that, while all other risk levels probability decreases with waiting time, the probability of No risk increases as waiting time increases.

Another major interest of this study is to visualize pedestrians risk at legal and illegal crossings. To create partial dependence plot, for this particular case, remaining variables would all be continuous, which if marginalized at their mean value would make the outcome less realistic and meaningful (shown in Figure 4.12). Therefore, for this and other following similar cases, a different approach of creating a combination of simulated

dataset of those continuous categorical variables has been taken. 10,000 random variables were generated in R from distribution of the numeric datatypes after identifying the distribution from their respective histogram. For example, the histogram of Wait_Time displayed exponentially decreasing trend, hence random values were generated from exponential distribution. All possible combinations of randomly generated Veh_gap and Ped_Speed values, with lanes, LE_ILLE, Veh_typ were merged to form a combined dataset. The combined dataset thus formed was used for creating the plot. The box plot shown in Figure 4.13 reveals the chances of "Fair Risk" and "High Risk" are notably higher in illegal crossings in comparison to legal crossings while the chance of "No Risk" being higher for the latter, as anticipated. Likewise, Figure 4.14 shows the probability density plot of the predicted probabilities from the same dataset.

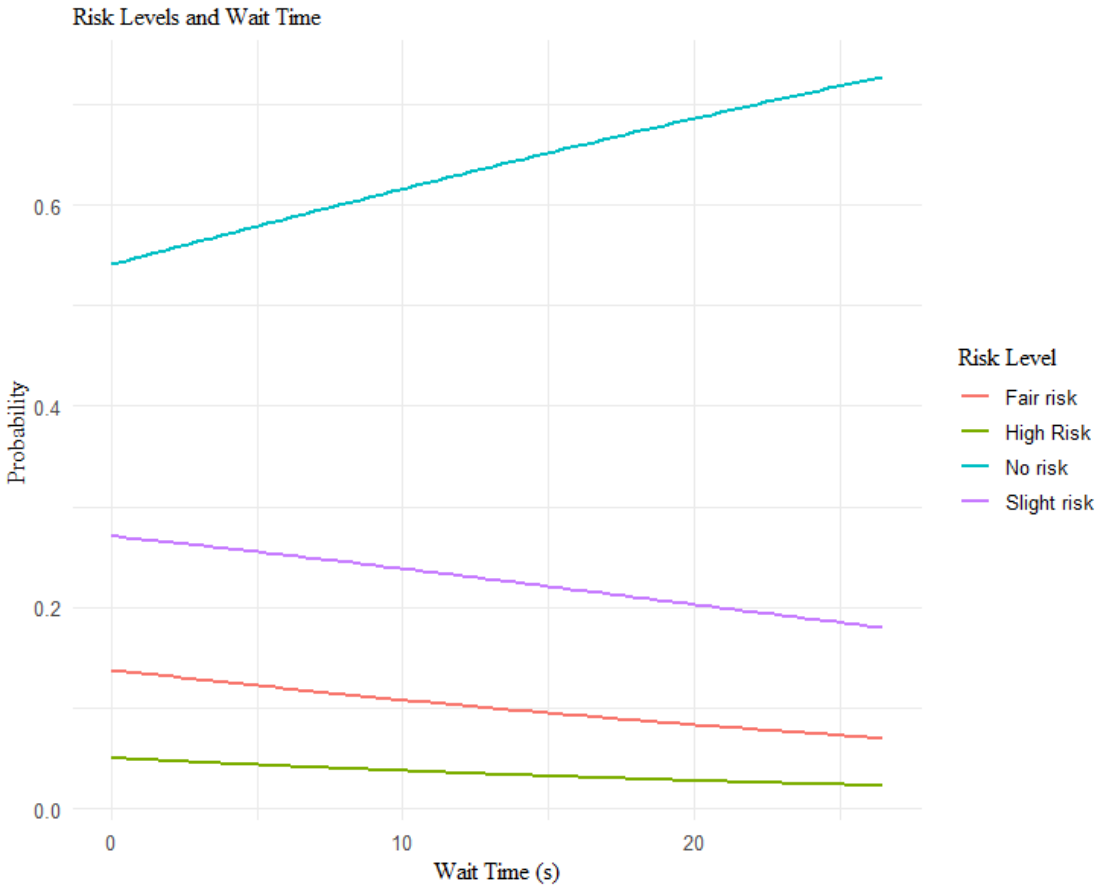


**Figure 4.11: PDP for Probability of conflict vs Wait_Time.**

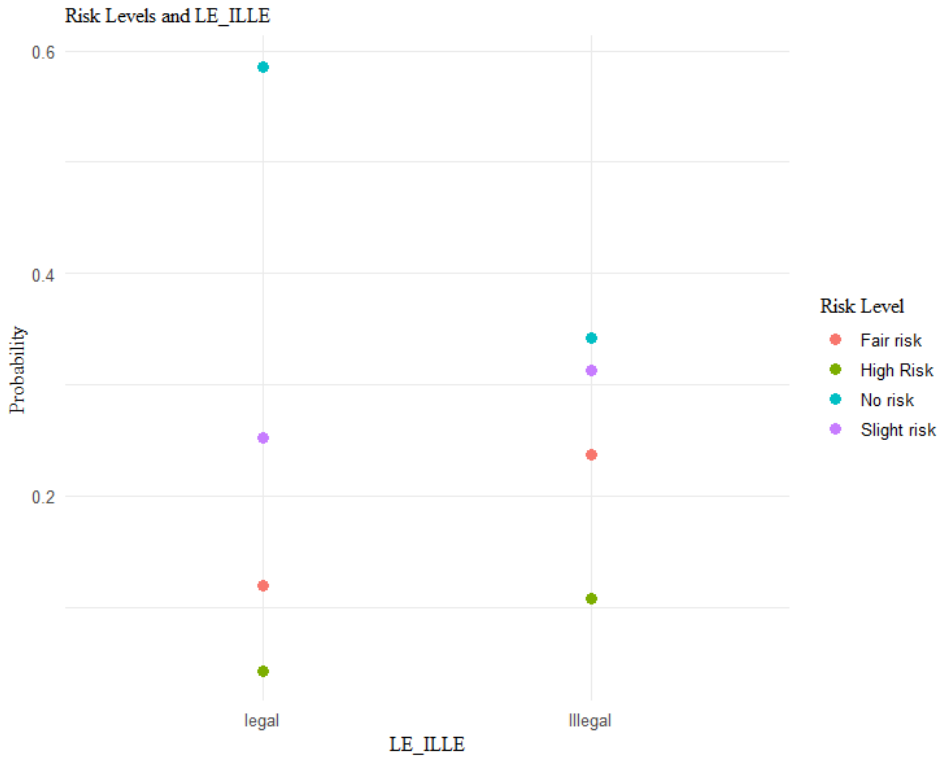


**Figure 4.12: PDP for Probability of conflict vs LE_ILLE**

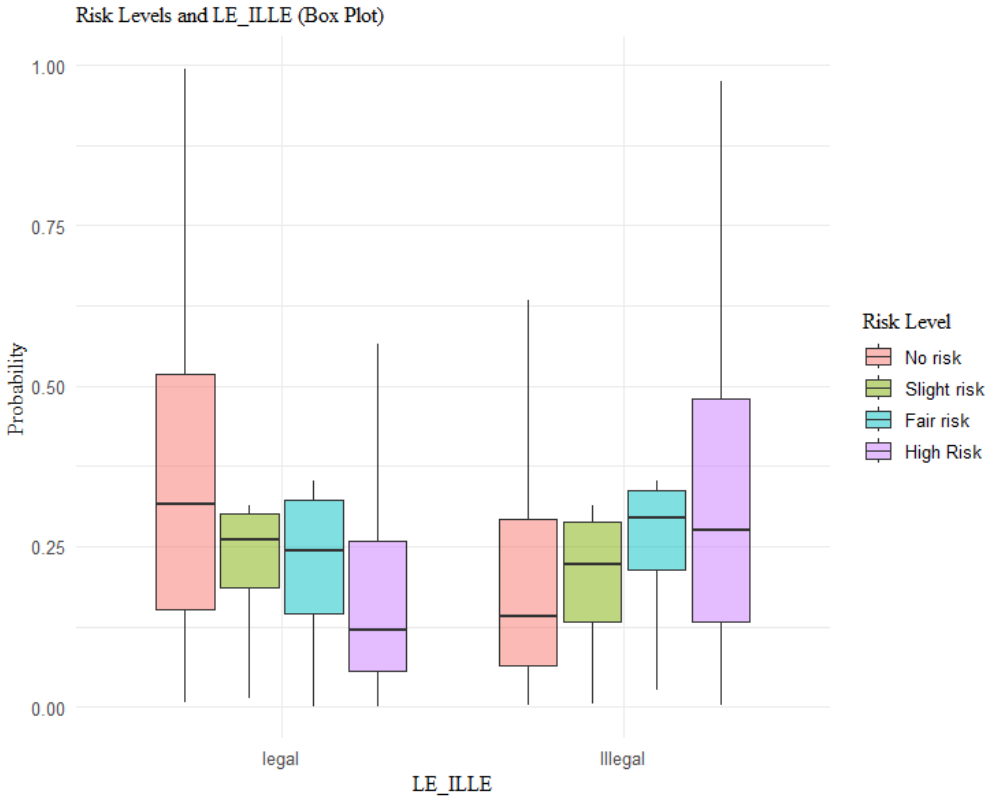


**Figure 4.13: Simulated boxplot for Probability of conflict vs LE_ILLE.**

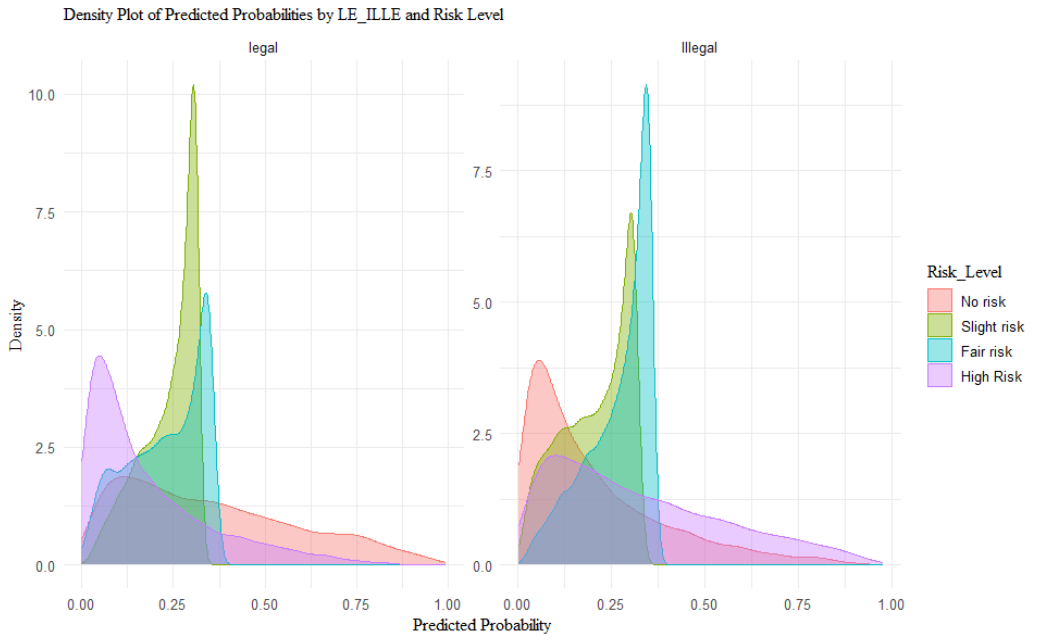


**Figure 4.14: Simulated Probability Density Plot for Probability of conflict vs LE_ILLE.**

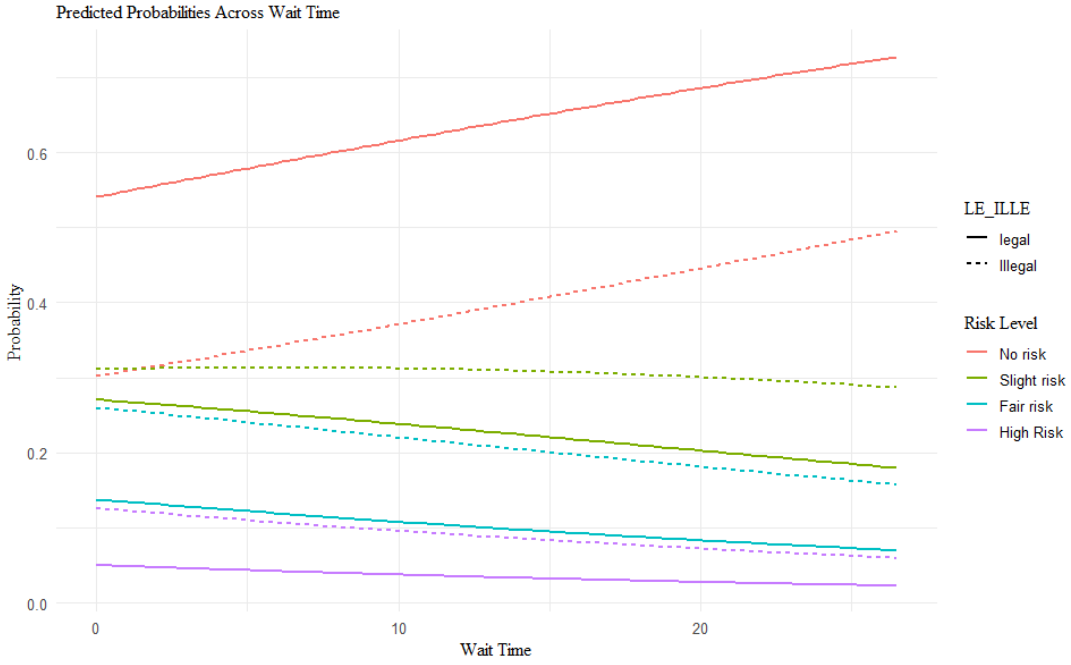


**Figure 4.15: PDP for Probability of conflict vs LE_ILLE and Wait_Time.**

To illustrate the variation of chances of various risk levels with Wait_Time, across legal and illegal crossings, partial dependence plot is depicted in Figure 4.15, which indicates

that the difference in chances of No risk in legal crossings is higher by constant value of nearly 0.2 for all wait times. Also, the chances of other levels of risk are lower in legal crossings and the plot for upper two levels of risk with respect to wait time can be seen more sensitive to illegal crossings while the opposite being true for "Slight Risk"(more sensitive for legal). Similar results can be seen for Pedestrian speed PDP plot (shown in Figure 4.16).One distinctive characteristic can be seen in one of the Figure 4.17 plots showing the variation of the Wait_Time plot across different age groups; specifically for age groups 20-40, probability of slight risk is found to be lesser for illegal crossings when the wait time is less than 5 secs.

Figure 4.18,4.19 and 4.20 illustrate the probabilities for three vehicle types across all four lanes at legal and illegal crossings. The first plot derived from the partial dependence method does not provide reliable output, for reasons mentioned above, as all other remaining variables are continuous. So, simulation method has been applied to generate latter two plots 4.19 (box plots) and 4.20 (plot for mean probabilities).The plots show the chances of "Fair Risk" and "High Risk" are higher at second and fourth lanes followed by third and first. Furthermore, it demonstrates that Two Wheelers pose higher risk in contrast to Car, Jeep, Van and heavy vehicles.

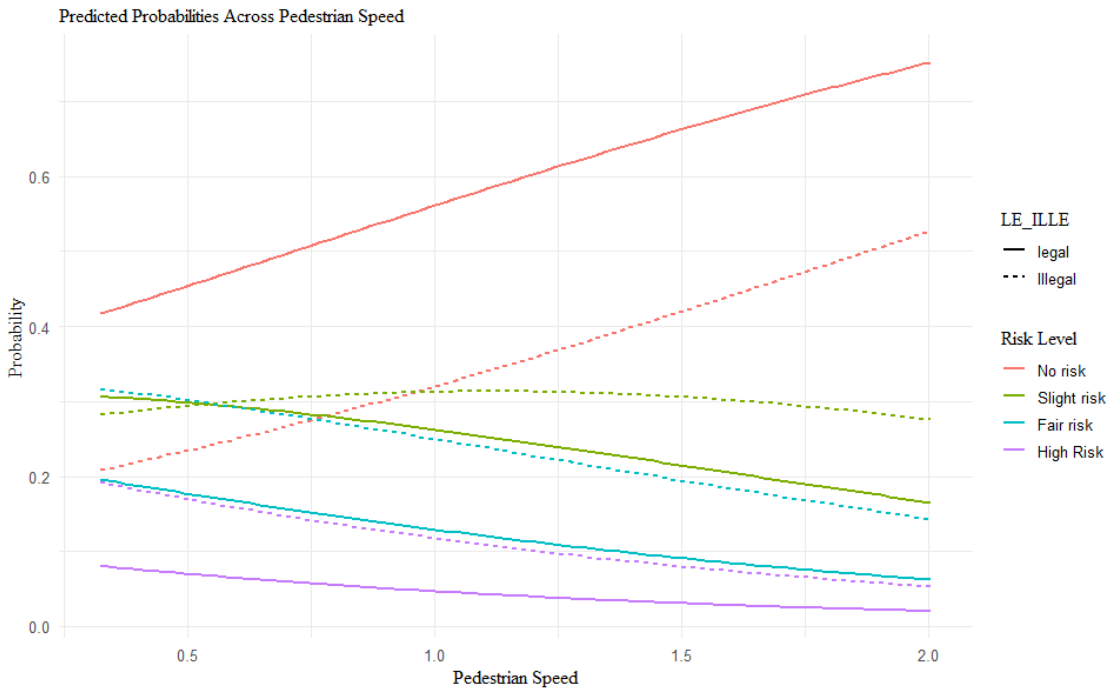


**Figure 4.16: PDP for Probability of conflict vs Ped_Speed and Wait_Time.**

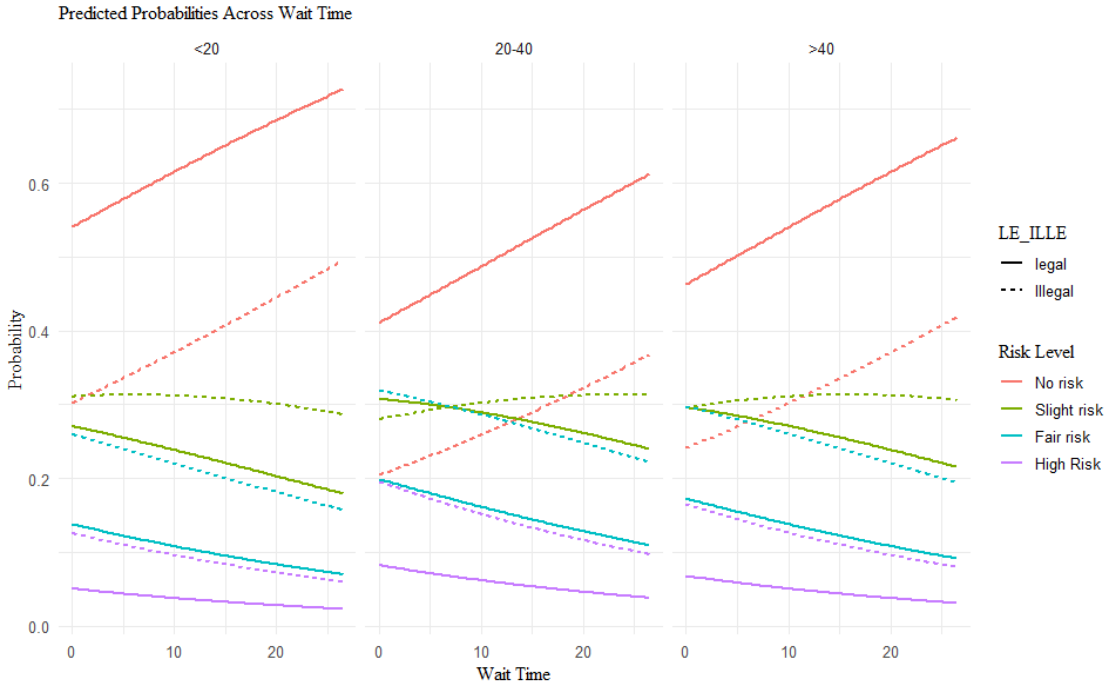


**Figure 4.17: PDP for Probability of conflict vs Age groups and Wait_Time.**

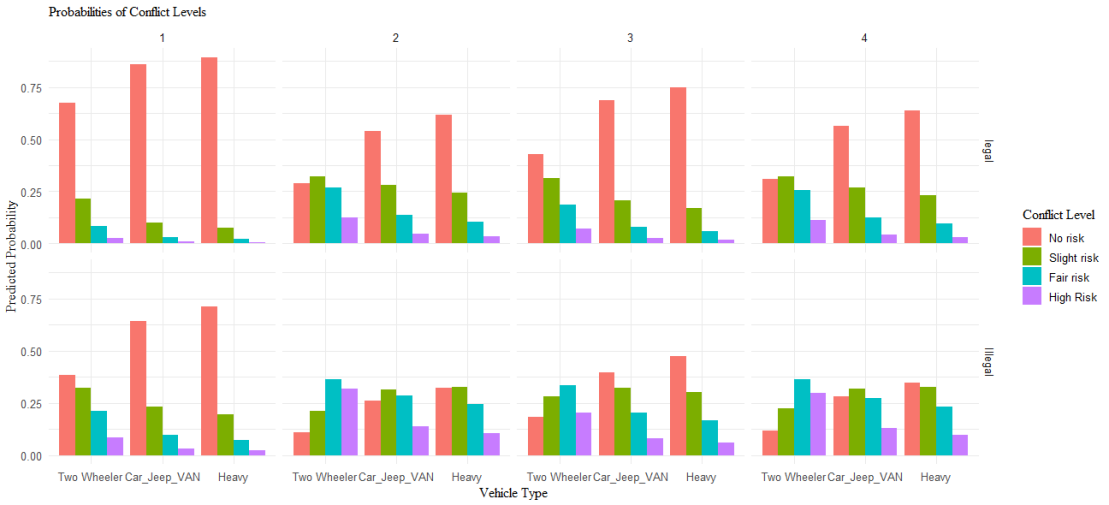


**Figure 4.18: PDP for Probability of conflict vs Lane, Vehicle Type and LE_ILLE.**

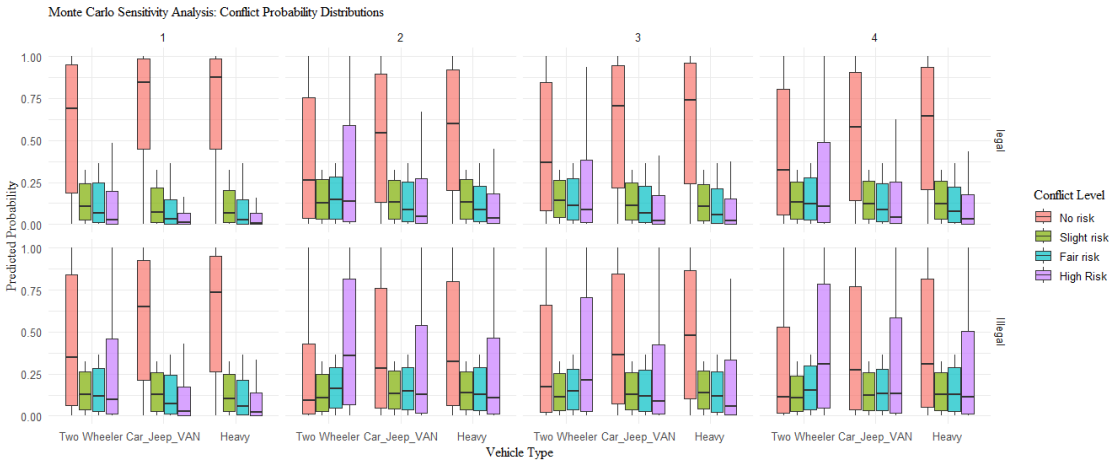


**Figure 4.19: Simulated boxplot for Probability of conflict vs Lane, Vehicle Type and LE_ILLE**


**Figure 4.20: Simulated Plot for Mean Probability of conflict vs Lane, Vehicle Type and LE_ILLE**

From Figure 4.21 and 4.22, it can be observed that the probability of "No Risk" increases and "High Risk" decreases with increasing accepted vehicular gaps for all cases, as expected. However, lines depicting "Slight Risk" and "Fair Risk" are seen to follow convex shape, peaking at some point of vehicular gap and then again decreasing. The inclination before peak indicates even though vehicular gaps are available to pedestrians, they are either crossing slowly or crossing at the end of the available vehicular gap which suggests at their inability to sense the risk or their willingness to endure those levels of risk. The peaks are shifting to the right as the vehicle size decreases. The shift could be because of pedestrians taking those risk (even when the accepted vehicular gaps is higher)

against smaller sizes vehicles by walking slowly ignoring the vehicles threat (perceiving the threat of bigger vehicles more) or from the vehicles perspective, smaller vehicles could be ignoring the safety of pedestrians, may be by not decelerating/stopping(decelerating/stopping before crossing increases available vehicle gap) as larger sizes vehicles do, to get past the pedestrians gap or accelerating before any immediate pedestrian platoon arrivals. This phenomenon is more evident in illegal crossings' plot.

Another simulation graph in Figure 4.23 elucidates similar feature for legal and illegal crossings; for lower vehicular gaps, pedestrians in legal crossings demonstrate more "Slight Risk" and "Fair Risk" taking behaviour while the probabilities of "High Risk" is prominent in later crossings.

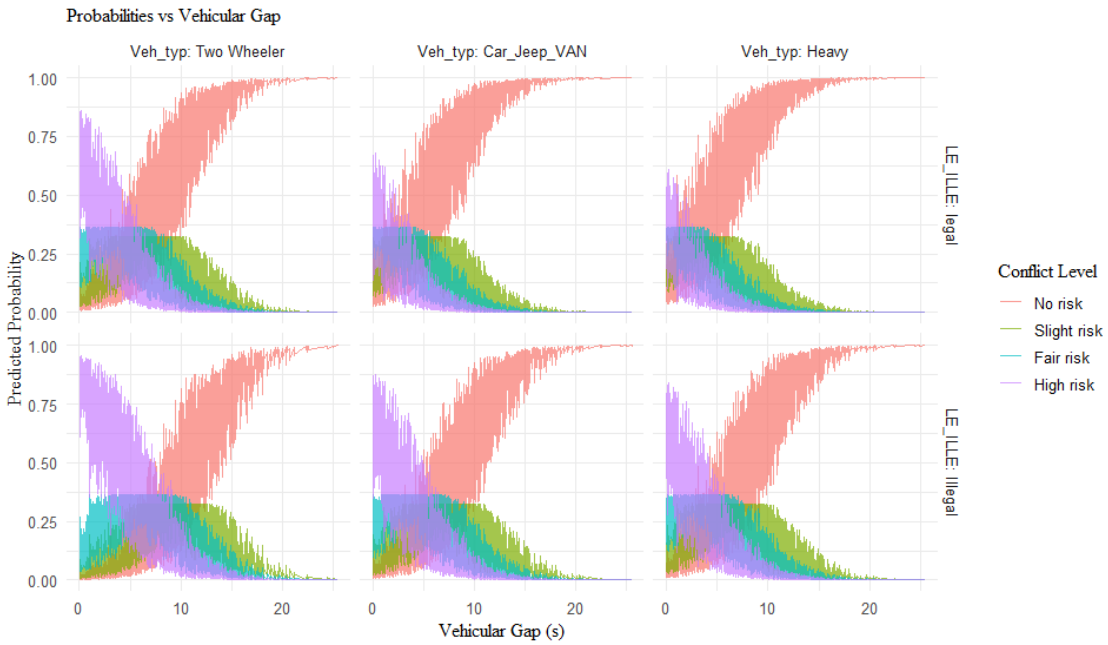


**Figure 4.21: Simulated plot for Probability of conflict vs Vehicular gap, LE_ILLE and Vehicle type.**

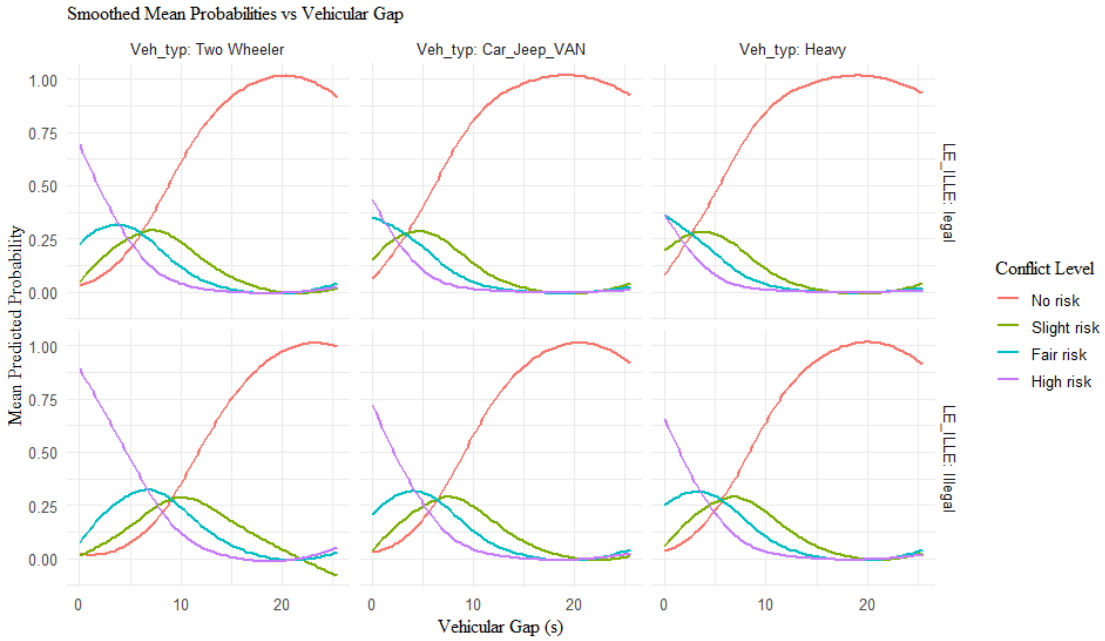


**Figure 4.22: Simulated best fit plot for Probability of conflict vs Vehicular gap, LE_ILLE and Vehicle type.**

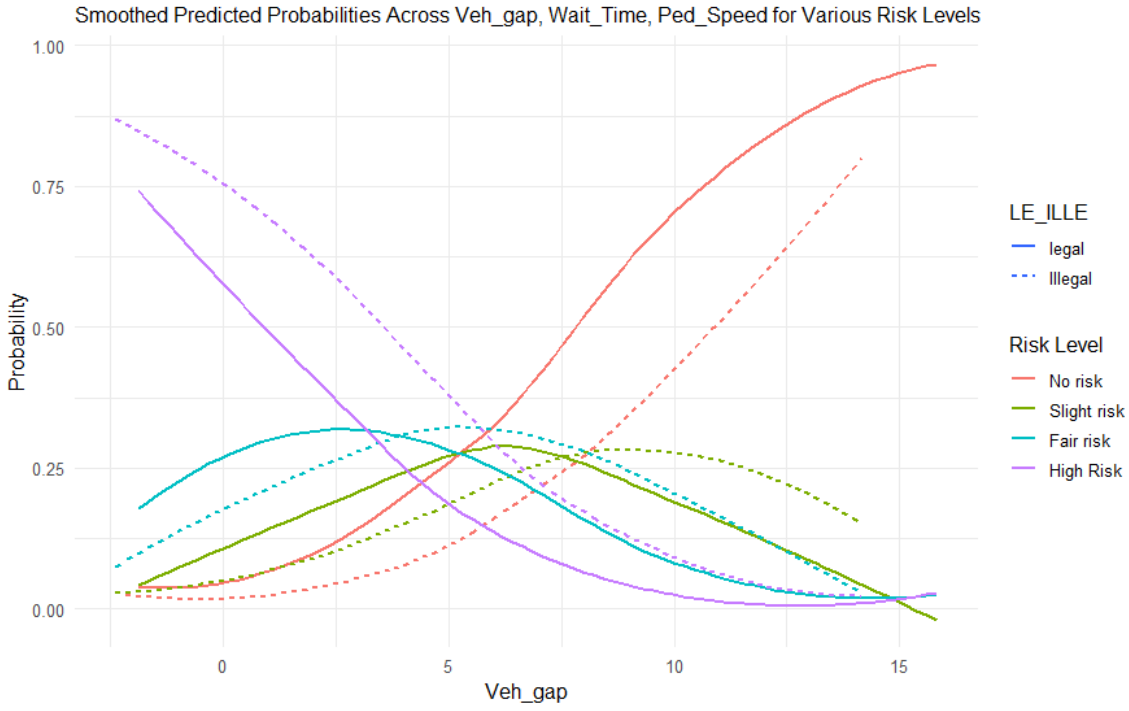


**Figure 4.23: PDP for Probability of conflict vs Vehicle gap and LE_ILLE**

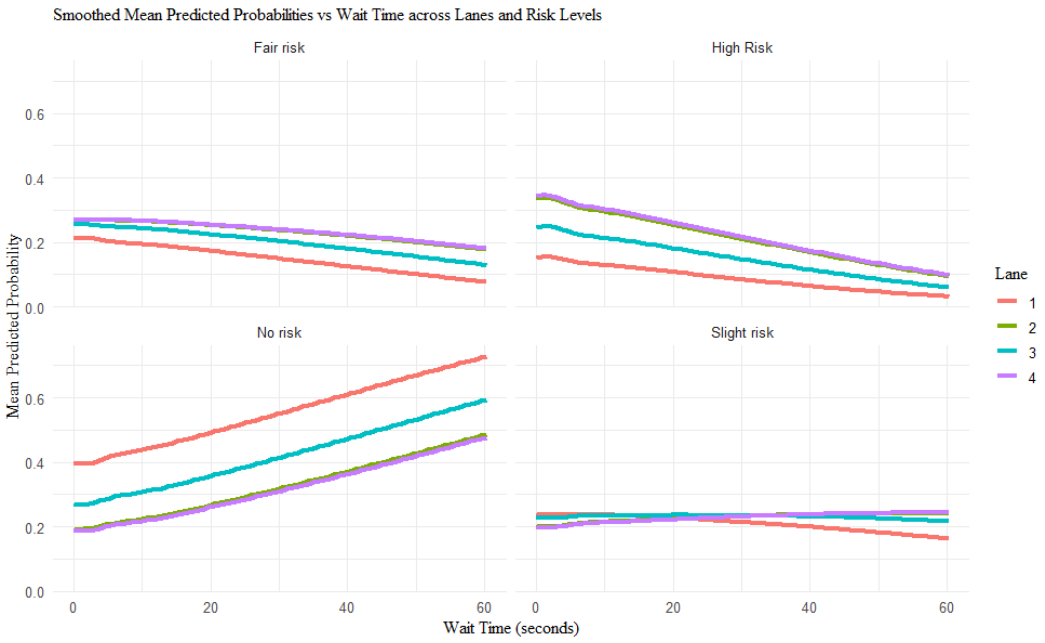


**Figure 4.24: Simulated plot for Probability of conflict vs Wait Time and lane.**

Figure 4.24 shows that while all other risk levels are sensitive to wait time, "Slight Risk" is comparatively independent. Moreover, "High Risk" probability is seen to decrease rapidly with increase wait time as compared to "Fair Risk".

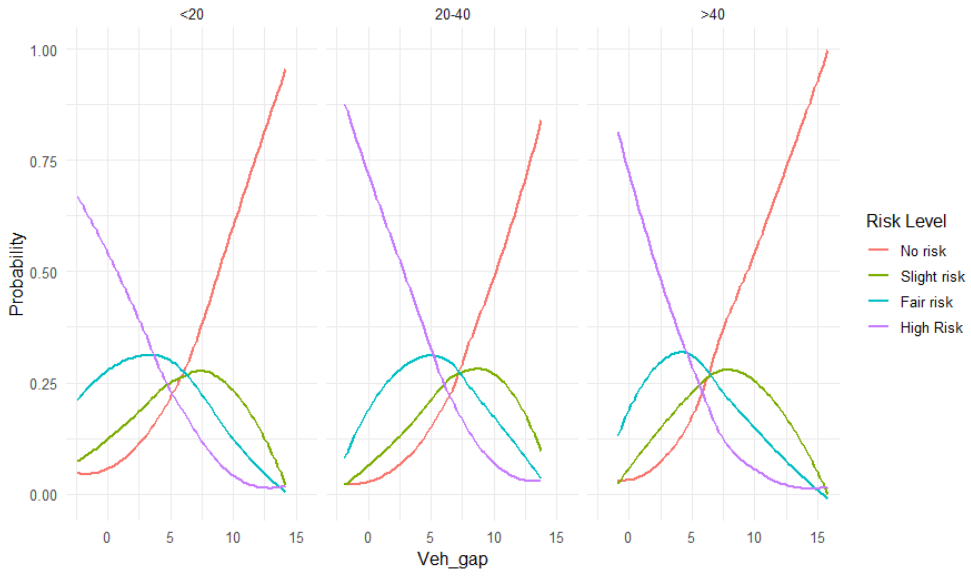


**Figure 4.25: Simulated plot for Probability of conflict vs Age and Vehicle gap.**

It can observed from Figure 4.25 that for similar vehicle gaps, pedestrians in age group 20-40 have higher chances of "High Risk" and besides, have the peaks of "Slight Risk"

and "Fair Risk" at larger vehicle gap values than the remaining two, suggesting those pedestrians in particular, have tendency for greater risk taking behavior.

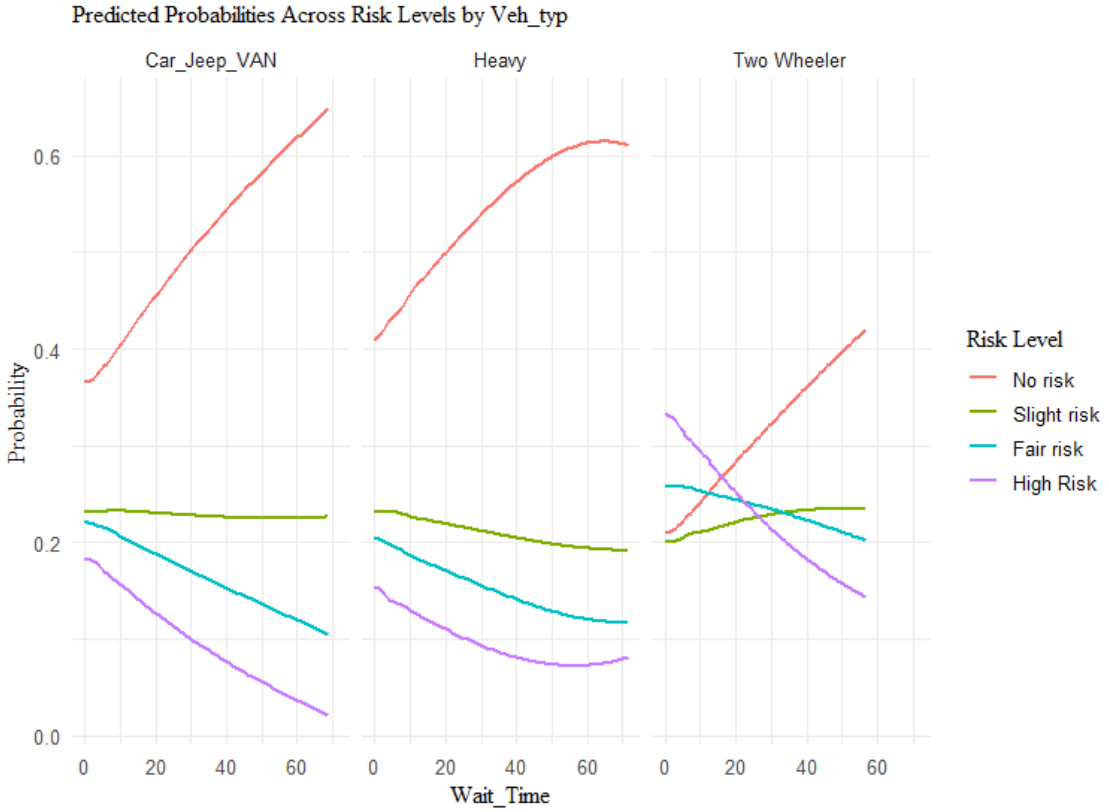


**Figure 4.26: Simulated plot for Probability of conflict vs Wait Time and Vehicle type**

One notable observation from simulated graph in 4.26 is that probability of high risk increases if the waiting time of pedestrians for heavy vehicles is more than 60 seconds. Also, the chances of slight risk against two wheelers increase as waiting time increases.

# CHAPTER 5 : CONCLUSION AND RECOMMENDATION

## 5.1 Conclusion

This research studied the factors affecting safety of pedestrians at legal and illegal crossings. To comprehend the effects of various variables explaining pedestrian characteristics/behavior, vehicle characteristics/behavior, and road features on safety margin, probability of conflict, and occurrence of four different levels of severity of conflict respectively, two different types of models-binary logit, and ordinal logit were developed. Results helped to derive the following important conclusions.

- The odds of severity increases by 1.21 in illegal crossings compared to legal ones with proper zebra and lane markings.
- The odds of conflict and its severity is maximum at second and fourth lanes respectively and since pedestrians waited longer at middle of the road (at the end of second lane), the odds is minimal at the third lane compared to the second and fourth.
- Pedestrians who spend more time at the road and the shelter (waiting time) are found to have less conflicting chances. Waiting time is negatively correlated with conflict by 0.05 and severity by 0.03.
- Increase in pedestrian speed is also found to decrease the odds of conflict by 1.93 and severity by 0.69.
- Both of the above conclusions suggest that pedestrians who cross the road with increased speed but spend longer time waiting at the beginning and at the end of lanes are more safer.
- Two wheelers are found to be riskiest; the odds decreased by 1.07 times for cars_jeep_van and by 1.38 for heavy vehicles like buses_trucks respectively. It might be because pedestrians perceived the threat of bigger vehicle size and kept more safety margin against heavy vehicles compared to smaller ones.
- Pedestrian while crossing in group tend to show more conflicting behaviour than crossing alone.

- Compared to younger age group of less than 20, pedestrians in age group 20-40 and greater than 40 are found to show more conflicting and severe risking behaviour.
- Acceptance of a unit large vehicular gaps is shown to decrease the odds of probability of conflict by 0.8 times and severity by 0.41 times.

## 5.2 Recommendation

The findings of this study can be utilized for evaluating pedestrian safety at midblock crossings and correspondingly develop appropriate crossing warrants based on pedestrian behaviour/demographics, vehicle characteristics/behaviour and road features. One notable observation regarding the relation of pedestrian speed and waiting time to odds of conflict and severity is that pedestrians who spend longer time on road but crosses the lanes with higher speed are safer than others. Moreover, two wheelers are found relatively riskier, so pedestrian have to be more alert while crossing against them. This study also highlights the importance of waiting shelters for pedestrians along with the risk they face while crossing multiple lanes. Further, research can be carried out considering more variables like pedestrian volume, vehicle flow rate, land use types and other pedestrian behaviour.

# REFERENCES


Chaudhari, A., Gore, N., Arkatkar, S., Joshi, G., & Pulugurtha, S. (2021). Exploring pedestrian surrogate safety measures by road geometry at midblock crosswalks: A perspective under mixed traffic conditions. *IATSS Research, 45*(1), 87–101. https://doi.org/10.1016/j.iatssr.2020.10.001

Chaudhari, A., Gore, N., Arkatkar, S., Joshi, G., & Parida, M. (2020). Choice crossing behaviour model for safety margin of pedestrians at mid-blocks in India. *Transportation Research Procedia, 48*, 2329–2342. https://doi.org/10.1016/j.trpro.2020.08.285

Chaudhari, A., Shah, J., Arkatkar, S., Joshi, G., & Parida, M. (2019). Evaluation of pedestrian safety margin at mid-block crosswalks in India. *Safety Science, 119*, 188–198. https://doi.org/10.1016/j.ssci.2018.12.009

Diogenes, M. C., & Lindau, L. A. (2010). Evaluation of pedestrian safety at midblock crossings, Porto Alegre, Brazil. *Transportation Research Record, 2193*(1), 37–43. https://doi.org/10.3141/2193-05

Golakiya, H., Patkar, M., & Dhamaniya, A. (2019). Analysis of vehicular pedestrian interaction at urban undesignated mid-block sections. In *Traffic safety and human behavior* (pp. 381–389). Springer. https://doi.org/10.1007/978-981-32-9042-6_31

Kadali, B. R., & Vedagiri, P. (2016). Proactive pedestrian safety evaluation at unprotected mid-block crosswalk locations under mixed traffic conditions. *Safety Science, 89*, 94–105. https://doi.org/10.1016/j.ssci.2016.06.017

Vedagiri, P., & Kadali, B. R. (2016). Evaluation of pedestrian–vehicle conflict severity at unprotected midblock crosswalks in India. Transportation Research Record, 2581(1), 48–56. https://doi.org/10.3141/2581-06

Khan, F. M., Jawaid, M., Chotani, H., & Luby, S. (1999). Pedestrian environment and behavior in Karachi, Pakistan. *Accident Analysis & Prevention, 31*(4), 335–339. https://doi.org/10.1016/S0001-4575(99)00002-2

Knoblauch, R. L., Pietrucha, M. T., & Nitzburg, M. (1996). Field studies of pedestrian walking speed and start-up time. *Transportation Research Record, 1538*, 27–38.

Kumar, J., & Suvash, S. (2010). Injuries and violence in Nepal: A review. *Injury Prevention, 16*(Suppl 1), A15. https://doi.org/10.1136/ip.2010.029215.540

Lobjois, R., & Cavallo, V. (2007). Age-related differences in street-crossing decisions: The effects of vehicle speed and time constraints on gap selection in an estimation task. *Accident Analysis & Prevention, 39*(5), 934–943. https://doi.org/10.1016/j.aap.2006.12.001

Louviere, J. J., Hensher, D. A., & Swait, J. D. (2000). *Stated choice methods: Analysis and application*. Cambridge University Press. https://doi.org/10.1017/CBO9780511753831.008

Ministry of Physical Infrastructure and Transport. (2013). *Nepal Road Safety Action Plan (2013-2020)*. Ministry of Physical Infrastructure and Transport. https://mopit.gov.np/en/sources/14/58756604

McFadden, D. (1974). *Conditional logit analysis of qualitative choice behavior*. In P. Zarembka (Ed.), *Frontiers in econometrics* (pp. 105–142). Academic Press.

Ojha, K. (2021). Road safety status and some initiatives in Nepal. *Journal of Engineering and Technology for Industrial Applications, 7*, 10. https://doi.org/10.5935/jetia.v7i27.713

Oxley, J., Fildes, B., Ihsen, E., Charlton, J., & Day, R. (1997). Differences in traffic judgments between young and old adult pedestrians. *Accident Analysis & Prevention, 29*(6), 839–847. https://doi.org/10.1016/S0001-4575(97)00052-1

Patel, M. R., Shukla, R. N., & Golakiya, H. D. (2018). Study of interaction between pedestrian and vehicle at undesignated urban mid-block section. *International Research Journal of Engineering and Technology, 5*(2), 56–72.

Tarawneh, M. S. (2001). Evaluation of pedestrian speed in Jordan with investigation of some contributing factors. *Journal of Safety Research, 32*(2), 229–236. https://doi.org/10.1016/S0022-4375(00)00058-5

Tiwari, G., Bangdiwala, S., Saraswat, A., & Gaurav, S. (2007). Survival analysis: Pedestrian risk exposure at signalized intersections. *Transportation Research Part F: Traffic Psychology and Behaviour, 10*(2), 77–89. https://doi.org/10.1016/j.trf.2006.08.002

Varhelyi, A. (1998). Drivers' speed behaviour at a zebra crossing: A case study. *Accident Analysis & Prevention, 30*(6), 731–743. https://doi.org/10.1016/S0001-4575(98)00015-2

Vedagiri, P., & Kadali, B. R. (2016). Evaluation of pedestrian–vehicle conflict severity at unprotected midblock crosswalks in India. *Transportation Research Record, 2581*(1), 48–56. https://doi.org/10.3141/2581-07

World Bank. (2020). *World Bank open knowledge repository*. https://openknowledge.worldbank.org/handle/10986/33324

Yannis, G., Papadimitriou, E., & Theofilatos, A. (2013). Pedestrian gap acceptance for mid-block street crossing. *Transportation Planning and Technology, 36*(5), 450–462. https://doi.org/10.1080/03081060.2013.818282

Zhang, C., Zhou, B., Chen, G., & Chen, F. (2017). Quantitative analysis of pedestrian safety at uncontrolled multi-lane mid-block crosswalks in China. *Accident Analysis & Prevention, 108*, 19–26. https://doi.org/10.1016/j.aap.2017.08.011

Zhang, Y., Yao, D., Qiu, T. Z., Peng, L., & Zhang, Y. (2012). Pedestrian safety analysis in mixed traffic conditions using video data. *IEEE Transactions on Intelligent*

*Transportation Systems*, *13*(4), 1832–1844. https://doi.org/10.1109/TITS.2012.2204066

Zhao, J., Malenje, J. O., Tang, Y., & Han, Y. (2019). Gap acceptance probability model for pedestrians at unsignalized mid-block crosswalks based on logistic regression. *Accident Analysis & Prevention*, *129*, 76–83. https://doi.org/10.1016/j.aap.2019.05.001

# APPENDIX A: Summary of Data

# APPENDIX B: Source Code for Binary Logit Model

```
library(openxlsx)
data<-
read.xlsx("C:/Users/asus/Desktop/analysis/midterm_data_analysis/fin_data_excel.xlsx",s
heet="final_data")
str(data)
summary(data)
library(dplyr)
nrow(data)

new_df <- data # here it was omit but later we remove na.omitt keep it below
summary(new_df)
str(new_df)
nrow(new_df)
class(new_df$Veh_typ)
str(new_df$Veh_typ)
summary(new_df$Veh_typ)
# to see what types of categorical variable are present
table(new_df$Veh_typ)
summary(new_df)
class(new_df)
str(new_df)
head(new_df)
#splitting into 4 (4 lanes) also while ordereing and factoring veh type only df1 (first
column was ordered and numerized so all columns must be numerized)
# also first of all na are removed from each of seaprated lane datas
df1<-new_df[,c(1:13)]
print(df1)
#omitting rows with cells of na value
fin_df1<-na.omit(df1)
head(df1)
df2<-new_df[,c(1:3,14:23)]
```

```
#omitting rows with cells of na value
fin_df2<-na.omit(df2)
head(fin_df2)
nrow(df2)
df3<-new_df[,c(1:3,24:33)]
print (df3)
#omitting rows with cells of na value
fin_df3<-na.omit(df3)
head(fin_df2)
df4<-new_df[,c(1:3,34:43)]
print (df4)
#omitting rows with cells of na value
fin_df4<-na.omit(df4)
head(fin_df4)
#conbine four to one dataset
four_df<-rbind(fin_df1,fin_df2,fin_df3,fin_df4)
print(four_df)
head(four_df)
print(nrow(four_df)
str(four_df)
summary(four_df)
#converting to factor
#see the distribution of Veh type and bin  in to classess
table(four_df$Veh_typ)
four_df$Veh_typ<-
factor(four_df$Veh_typ,levels=c("CYCLE","SCOOTER","M","CAR","C","JEEP","SCO
RPIO","VAN","V","DELIVERY
VAN","B","H","HEAVY","CRANE","TRUCK"),labels=c("Two Wheeler","Two
Wheeler","Two
Wheeler","Car_Jeep_VAN","Car_Jeep_VAN","Car_Jeep_VAN","Car_Jeep_VAN","Car
_Jeep_VAN","Car_Jeep_VAN","Car_Jeep_VAN","Heavy","Heavy","Heavy","Heavy","
Heavy"))
table(four_df$Veh_typ)
four_df$Age<-factor(four_df$Age,levels=c("1","2","3"),labels=c("<20","20-40",">40"))
```

```
table(four_df$Age)

four_df$Gen<-as.factor(four_df$Gen)
table(four_df$Gen)
four_df$Dry<-as.factor(four_df$Dry)
four_df$LE_ILLE<-
factor(four_df$LE_ILLE,levels=c("0","1"),labels=c("Legal","Illegal"))
table(four_df$LE_ILLE)

four_df$lane<-as.factor(four_df$lane)
table(four_df$lane)
#grouping Gr_size converting in to factor,# based on equal class conditioning# for linear
modeelinf gr size is kept numeric
#see the distribution of Gr_Size
table(four_df$GR_Size)
str(four_df$GR_Size)
four_df$GR_Size<-cut(four_df$GR_Size,breaks=c(0,1,2.99,4.99,8),labels=c("1","2" ,"(3-
4)","(5-7)" )) # (0,1], (1,3],(3,7] based on class distribution
str(four_df$GR_Size)
str(four_df)

#summary of data
library(dplyr)
library(gtsummary)
library(tidyverse)
tbl<-tbl_summary(four_df)
tbl
# Convert tbl_summary object to data frame
tbl_df <- as.data.frame(tbl)
# Write the data frame to a CSV file
write.csv(tbl_df, file = "write.csv", row.names = FALSE)

#correlation between numeric variables
#filter numervc valyues from data set
```

```
cormat1<-cor(four_df[sapply(four_df,is.numeric)])
cormat1
#exclude some data from correlation matrixPTC and SM
cormat<-cormat1[-c(4,6,8),-c(4,6,8)]
round(cormat,2)
#visualize correlation matrix
library(corrplot)
corrplot(cormat,method="number")
str(four_df)
#check from other method
cor(four_df$Ped_Speed,four_df$PTC)
#corerlation betwen categorical variables not good so cramers V used
#chisq.test(table(four_df$Age, four_df$Gen))

#correlation method categorical variables 2 cramers V
 # creaion of datafrma with only categorical varibales
df_cramer<- four_df[sapply(four_df,is.factor)]
head(df_cramer)
summary(df_cramer)

#cramers V test apply chat gpt code
Copy code
# Load the vcd package
library(vcd)
# Create a list to store the results
cramers_v_results <- list()
# Loop through all combinations of variables in df_cramer and df
for (i in 1:(ncol(df_cramer))) {
  for (j in 1:(ncol(df_cramer))) {
    # Create contingency table
    cont_table <- table(df_cramer[,i], df_cramer[,j])

    # Compute Cramer's V
    cramers_v <- assocstats(cont_table)$cramer
```

```
    # Store the result
    cramers_v_results[[paste(names(df_cramer)[i], names(df_cramer)[j], sep="_vs_")]] <-
cramers_v
  }
}
# Print the results
print(cramers_v_results)

#draw heat map of cramers V results from excel table# cramers V results re manually
enetred in neatatb file , sorrelation _categorical sheet and then this code is used to plot
heat maps
library(openxlsx)
cramer_heat <-
read.xlsx("C:/Users/asus/Desktop/analysis/midterm_data_analysis/correlation.xlsx",sheet
="correlation_categorical")
cramer_heat
cramer_matheat<-data.matrix(cramer_heat[,2:7])
rownames(cramer_matheat) <- c("Age","Gen","Veh_typ","lane","LE_ILLE","Dry")
cramer_matheat
class(cramer_matheat)
library(corrplot)
corrplot(cramer_matheat,method="number")
# for logit model with out data cleaning (interquartile range)
boxplot(four_df)$out
summary(four_df)
# add new column of conflict where sm if <1,-1 if >1 0
data01 <- transform(four_df, conflict= ifelse(SM>1, 0, 1))
table(data01$SM)
summary(data01)
head(data01)
#check class of data
class(data01$SM)
class(data01$conflict)
```

```
#convert conflict to factor variables (after multiple trial it was seen that this conversion
was not necessary)
data01$conflict=as.factor(data01$conflict)
table(data01$conflict)
#splitting data into training and testing set
library(caTools)
#make this example reproducible
set.seed(123)
#use 80% of dataset as training set and 20% as test set
sample <- sample.split(data01$conflict, SplitRatio = 0.8)
train  <- subset(data01, sample == TRUE)
test   <- subset(data01, sample == FALSE)
# see no of rows in df train and dftest
nrow(train)
nrow(test)
#plot graphs to visualize logit model before fitting curve
library(ggplot2)
ggplot(data01,aes(x=Wait_Time,y=as.factor(conflict)))+  geom_jitter(height=.05,alpha=.1)
#logit model in R
model <- glm(conflict ~
LE_ILLE+Gen+Veh_gap+lane+Age+Veh_typ+GR_Size+Dry+Ped_Speed+Veh_Speed+
Wait_Time,family=binomial(link='logit'),data=train)
summary(model)
#removal of non significant variables
model <- glm(conflict ~
Veh_gap+Age+GR_Size+Dry+Ped_Speed+Veh_Speed,family=binomial(link='logit'),dat
a=train)
summary(model)
# to change reference class
# make Crew the reference group in Class not needed (run before train split)
#data01$Veh_typ = relevel(data01$Veh_typ, ref = "Car_JS")

# SINCE MANY VALUES ARE INSIGNIFICANT THERE ARE THREE
IDEAS( INCREASE SAMPLE SIZE, STEPWISE,REMOVE OUTLIERS)
```

```
#REMOVE OUTLIERS
str(data01)
#gen has three lavels instead of two
# checking where there is "M " M with a space instead of M data01[data01$Gen=="M ",]
#table(data01$Gen)

#removin outliers method 1
boxplot(data01)$out
summary(data01)
library(ggstatsplot)
library(ggplot2)
library(rstantools)
Q <- quantile(data01$Wait_Time, probs=c(.25, .75), na.rm = FALSE)
iqr <- IQR(data01$Wait_Time)
up <-  Q[2]+1.5*iqr # Upper Range
low<- Q[1]-1.5*iqr # Lower Range
eliminated1<- subset(data01, data01$Wait_Time > (Q[1] - 1.5*iqr) & data01$Wait_Time
< (Q[2]+1.5*iqr))
summary(eliminated1)
boxplot(eliminated1)
Q <- quantile(data01$Wait_Time, probs=c(.25, .75), na.rm = FALSE)
iqr <- IQR(data01$Wait_Time)
up <-  Q[2]+1.5*iqr # Upper Range
low<- Q[1]-1.5*iqr # Lower Range
eliminated1<- subset(eliminated1, eliminated1$Wait_Time > (Q[1] - 1.5*iqr) &
eliminated1$Wait_Time < (Q[2]+1.5*iqr))
summary(eliminated1)
boxplot(eliminated1)

Q <- quantile(eliminated1$Veh_gap, probs=c(.25, .75), na.rm = FALSE)
iqr <- IQR(eliminated1$Veh_gap)
up <-  Q[2]+1.5*iqr # Upper Range
low<- Q[1]-1.5*iqr # Lower Range
```

```
eliminated<- subset(eliminated1, eliminated1$Veh_gap > (Q[1] - 1.5*iqr) &
eliminated1$Veh_gap < (Q[2]+1.5*iqr))
boxplot(eliminated)
str(eliminated)
Q <- quantile(eliminated$SM, probs=c(.25, .75), na.rm = FALSE)
iqr <- IQR(eliminated$SM)
up <-  Q[2]+1.5*iqr # Upper Range
low<- Q[1]-1.5*iqr # Lower Range
eliminated2<- subset(eliminated, eliminated$SM > (Q[1] - 1.5*iqr) & eliminated$SM <
(Q[2]+1.5*iqr))
boxplot(eliminated2)
Q <- quantile(eliminated2$Ped_Speed, probs=c(.25, .75), na.rm = FALSE)
iqr <- IQR(eliminated2$Ped_Speed)
up <-  Q[2]+1.5*iqr # Upper Range
low<- Q[1]-1.5*iqr # Lower Range
eliminated3<- subset(eliminated2, eliminated2$Ped_Speed > (Q[1] - 1.5*iqr) &
eliminated2$Ped_Speed < (Q[2]+1.5*iqr))
boxplot(eliminated3)

Q <- quantile(eliminated3$Veh_Speed, probs=c(.25, .75), na.rm = FALSE)
iqr <- IQR(eliminated34$Veh_Speed)
up <-  Q[2]+1.5*iqr # Upper Range
low<- Q[1]-1.5*iqr # Lower Range
eliminated4<- subset(eliminated3, eliminated3$Veh_Speed > (Q[1] - 1.5*iqr) &
eliminated3$Veh_Speed < (Q[2]+1.5*iqr))
boxplot(eliminated4)
summary(eliminated4$Wait_Time)

# START OF BINARY LOGIT MODEL making eliminated as data01 as all below has
data01
data01<-eliminated4
table(data01$conflict)
#for gen extra M
table(data01$Gen)
```

```
# Find row numbers for each category in the 'Category' column
category_rows <- lapply(unique(data01$Gen), function(cat) which(data01$Gen == cat))
# Output
names(category_rows) <- unique(data01$Gen)
print(category_rows)
#actual vehicle speed (speed*dry)#deemed not good
#data01$vehactsped<-data01$Veh_Speed*(as.numeric(data01$Dry))
#head(data01)

#change reference category #no any effect so remove
#data01$Age=relevel(data01$Age,ref=">40")
#data01$GR_Size=relevel(data01$GR_Size,ref="(5-7)")
#data01$Veh_typ=relevel(data01$Veh_typ,ref="Heavy")
#making logit model after removing outliers (eliminated is final dataset)
#splitting data into training and testing set
library(caTools)
#make this example reproducible
set.seed(123)
#use 80% of dataset as training set and 20% as test set
sample <- sample.split(data01$conflict, SplitRatio = 0.75)
train  <- subset(data01, sample == TRUE)
test   <- subset(data01, sample == FALSE)
# see no of rows in df train and dftest
nrow(train)
nrow(test)

#plot graphs to visualize logit model before fitting curve
library(ggplot2)
ggplot(data01,aes(x=Wait_Time,y=conflict))+ geom_jitter(height=.05,alpha=.1)
#logit model in R
model <- glm(conflict
~GR_Size+lane+Veh_gap+Ped_Speed+Wait_Time,family=binomial(link='logit'),data=tr
ain)
summary(model)
```

```
model2 <- glm(conflict
~Age+GR_Size+lane+Veh_gap+Ped_Speed+Dry,family=binomial(link='logit'),data=trai
n)
summary(model2)
model_all <- glm(conflict
~Veh_Speed+lane+Veh_gap+Ped_Speed+Wait_Time+Gen+Veh_typ+GR_Size+LE_ILL
E+Dry+Age,family=binomial(link='logit'),data=train)
summary(model_all)
#Stepwise regression in R backward stepwisemthod 1
model <- step(model_all, direction = "backward")

#Stepwise regression in R both dierction stepwisemthod 2
#model is model with all variables
both_model <- step(model_all, direction = "both")
summary(both_model)
#same result from forward and backward stepwise regression conflict ~ Age + lane +
Veh_gap + Ped_Speed + GR_Size
#keep wait time

#removing insignifiant varaibels
model <- glm(conflict
~Age+Gen+Veh_gap+Ped_Speed,family=binomial(link='logit'),data=train)
summary(model)
####################sensiivity analysis for lane,ped_speed############
str(data01)
# Define a grid for Ped_Speed
ped_speed_grid <- seq(min(data01$Ped_Speed, na.rm = TRUE),
                      max(data01$Ped_Speed, na.rm = TRUE),
                      length.out = 100)
# Set other variables to fixed values
fixed_values <- data.frame(
  GR_Size = "1",  # Replace with the desired reference category for analysis
  lane = c("1", "2", "3", "4"), # Include all lanes
  Age = "20-40",
```

```
  Gen = "F",
  Veh_gap = mean(data01$Veh_gap, na.rm = TRUE),
  Wait_Time = mean(data01$Wait_Time, na.rm = TRUE)

)
# Create a data frame to hold all combinations of Ped_Speed and lane
sensitivity_data <- expand.grid(
  Ped_Speed = ped_speed_grid,
  lane = fixed_values$lane
)
# Ensure factor levels are correctly matched
sensitivity_data$lane <- factor(rep(fixed_values$lane, each = length(ped_speed_grid)),
levels = levels(data01$lane))
sensitivity_data$Gen <- factor(fixed_values$Gen)
sensitivity_data$Age <- factor(fixed_values$Age)
# Add fixed values for all predictors
sensitivity_data$GR_Size <- factor(fixed_values$GR_Size,
                                   levels = c("1", "2", "(3-4)", "(5-7)"))
sensitivity_data$Veh_gap <- fixed_values$Veh_gap
sensitivity_data$Wait_Time <- fixed_values$Wait_Time
# Predict probabilities using the logistic regression model
sensitivity_data$predicted_prob <- predict(model,
                                           newdata = sensitivity_data,
                                           type = "response")
# Load necessary library
library(ggplot2)
# Plot the sensitivity graph
####################sensiivity analysis for lane,wait_time############
str(data01)
# Define a grid for Wait_Time
Wait_Time_grid <- seq(min(data01$Wait_Time, na.rm = TRUE),
                      max(data01$Wait_Time, na.rm = TRUE),
                      length.out = 100)
# Set other variables to fixed values
```

```
fixed_values <- data.frame(
  GR_Size = "1",  # Replace with the desired reference category for analysis
  lane = c("1", "2", "3", "4"), # Include all lanes
  Age = ">40",
  Gen = "F",
  Veh_gap = mean(data01$Veh_gap, na.rm = TRUE),
  Ped_Speed = mean(data01$Ped_Speed, na.rm = TRUE)


)
# Create a data frame to hold all combinations of Ped_Speed and lane
sensitivity_data <- expand.grid(
  Wait_Time = Wait_Time_grid,
  lane = fixed_values$lane
)
# Ensure factor levels are correctly matched
sensitivity_data$lane <- factor(rep(fixed_values$lane, each = length(ped_speed_grid)),
levels = levels(data01$lane))
sensitivity_data$Gen <- factor(fixed_values$Gen)
sensitivity_data$Age <- factor(fixed_values$Age)
# Add fixed values for all predictors
sensitivity_data$GR_Size <- factor(fixed_values$GR_Size,
                                   levels = c("1", "2", "(3-4)", "(5-7)"))
sensitivity_data$Veh_gap <- fixed_values$Veh_gap
sensitivity_data$Ped_Speed <- fixed_values$Ped_Speed
# Predict probabilities using the logistic regression model
sensitivity_data$predicted_prob <- predict(model,
                                           newdata = sensitivity_data,
                                           type = "response")
# Load necessary library
library(ggplot2)
# Plot the sensitivity graph
ggplot(sensitivity_data, aes(x = Wait_Time, y = predicted_prob, color = lane, group =
lane)) +
  geom_line(size = 1) +
```

```
  labs(
    title = "Predicted Probability vs Lane and Wait_Time",
    x = "Wait_Time (m/s)",
    y = "Predicted Probability",
    color = "Lane"
  ) +
  theme_minimal(base_size = 12) +
  theme(
    plot.title = element_text(family = "Times", size = 12, color = "black"),
    legend.title = element_text(family = "Times", size = 12, color = "black"),  # Try a
default font
    axis.title.x = element_text(family = "Times", size = 12, color = "black"),  # X-axis
label font
    axis.title.y = element_text(family = "Times", size = 12, color = "black"),
  )

#####3d plot#####
library(plotly)
library(reshape2)
sensitivity_matrix <- acast(sensitivity_data, Ped_Speed ~ Wait_Time, value.var =
"predicted_prob")
# Generate 3D Surface Plot
fig <- plot_ly(
  x = ped_speed_grid,
  y = wait_time_grid,
  z = sensitivity_matrix,
  type = "surface"
)
# Add labels and title
fig <- fig %>%
  layout(
    title = "3D Sensitivity Analysis: Predicted Probability vs. Pedestrian Speed & Wait
Time",
 legend = list(
```

```
    title = list(text = "Cylinders", font = list(family = "Times New Roman", size = 12,
color = "black"))
   ),

scene = list(
    xaxis = list(title = "Pedestrian Speed (m/s)"),
    yaxis = list(title = "Wait Time (s)"),
    zaxis = list(title = "Predicted Probability")
   )
 )
# Show the plot
fig
#######

############GR_Size and Wait Time ignore negative waittime values#####

# Load necessary libraries
library(ggplot2)
# Define grid for GR_Size (categorical variable)
gr_size_levels <- c("1", "2", "(3-4)", "(5-7)")
# Define grid for Wait_Time, ignoring negative values
wait_time_grid <- seq(min(data01$Wait_Time[data01$Wait_Time >= 0], na.rm = TRUE),
                      max(data01$Wait_Time, na.rm = TRUE),
                      length.out = 100)
# Create fixed values for other variables
fixed_values <- data.frame(
  Ped_Speed = mean(data01$Ped_Speed, na.rm = TRUE),
  Age = "<20",
  Gen = "M",
  lane = "1",  # Specify a single fixed value for lane
  Veh_Speed = mean(data01$Veh_Speed, na.rm = TRUE),
  Veh_gap = mean(data01$Veh_gap, na.rm = TRUE)
)
# Create a data frame to hold all combinations of GR_Size and Wait_Time
```

```
sensitivity_data <- expand.grid(
  GR_Size = gr_size_levels,
  Wait_Time = wait_time_grid
)
# Convert GR_Size to a factor with correct levels
sensitivity_data$GR_Size <- factor(sensitivity_data$GR_Size,
                                   levels = c("1", "2", "(3-4)", "(5-7)"))
# Add fixed values for other variables
sensitivity_data$Ped_Speed <- fixed_values$Ped_Speed
sensitivity_data$Age <- factor(fixed_values$Age)
sensitivity_data$Gen <- factor(fixed_values$Gen)
sensitivity_data$lane <- factor(fixed_values$lane, levels = levels(data01$lane))
sensitivity_data$Veh_Speed <- fixed_values$Veh_Speed
sensitivity_data$Veh_gap <- fixed_values$Veh_gap
# Predict probabilities using the logistic regression model
sensitivity_data$predicted_prob <- predict(model,
                                           newdata = sensitivity_data,
                                           type = "response")
# Plot: GR_Size vs. Predicted Probability for different Wait_Time values
ggplot(sensitivity_data, aes(x = Wait_Time, y = predicted_prob, color = GR_Size, group
= GR_Size)) +
  geom_line(size = 1) +
  scale_color_manual(values = c("blue", "green", "orange", "red")) +  # Custom colors for
GR_Size
  labs(
    title = "Predicted Probability vs. Group Size and Wait Time",
    x = "Wait Time (s)",
    y = "Predicted Probability",
    color = "Group Size"
  ) +
  theme_minimal(base_size = 12) +
  theme(
    plot.title = element_text(family = "Times", size = 12, color = "black"),
    legend.title = element_text(family = "Times", size = 12, color = "black"),
```

```
    axis.title.x = element_text(family = "Times", size = 12, color = "black"),
    axis.title.y = element_text(family = "Times", size = 12, color = "black")
  )

#prediction https://www.r-bloggers.com/2015/09/how-to-perform-a-logistic-regression-in-r/
fitted.results <- predict(model,newdata=subset(test),type='response')
fitted.results <- ifelse(fitted.results > 0.5,1,0)
print(fitted.results)
misClasificError <- mean(fitted.results != test$conflict)
print(paste('Accuracy',1-misClasificError))
#ROC curve
library(ROCR)
p <- predict(model, newdata=test, type="response")
pr <- prediction(p, test$conflict)
prf <- performance(pr, measure = "tpr", x.measure = "fpr")
plot(prf)
auc <- performance(pr, measure = "auc")
auc <- auc@y.values[[1]]
auc
# Save the plot to a PNG file
png("prf.png", width = 800, height = 600, units = "px", res = 100)
plot(prf)
dev.off()
# pseudo square
library(pscl)
pR2(model)
#confusion matrix method 1
print(test$conflict)
print(fitted.results)
pred1<-as.factor(fitted.results)
test$conflict<-as.factor(test$conflict)
class(pred1)
library(caret)
```

```
example <- confusionMatrix(data=pred1, reference = test$conflict)
example
#confusion matrix method 2
table1 <-table(Predicted=fitted.results, Actual=test$conflict)
print(table1)
confusionMatrix(table1)
#END OF BINARY LOGIT MODEL
```

# APPENDIX C: Source Code for Ordinal Logit Model

```
#RI cluster addition of RI columns
str(eliminated4)
table(eliminated4$RI)
#rescaline SM because of near 0 SM values
#library(scales)
#eliminated4$rescaledSM <- rescale(eliminated4$SM, to = c(1, 2))
#str(eliminated4)
#summary(vec_range4)
# now RI=speed/SM
# Find the minimum value in the range
min_value <- min(eliminated4$SM)
# Shift the entire range upwards by adding the absolute value of the minimum value
eliminated4$scaledSM <- eliminated4$SM + abs(min_value)
str(eliminated4$scaledSM)
summary(eliminated4$scaledSM)
eliminated4$RI<-(eliminated4$Veh_Speed/eliminated4$scaledSM)
eliminated4$logRI<-log(eliminated4$RI)
str(eliminated4$RI)
summary(eliminated4$RI)
str(eliminated4$logRI)
summary(eliminated4$logRI)
summary(eliminated4)
table(eliminated4$RI)
#replace infinite values with NA
eliminated4$RI[is.infinite(eliminated4$RI)] <- NA
# calculate CDF
CDF <- ecdf(eliminated4$RI)
# draw the cdf plot
plot( CDF )
#finding the values at 25%,40% and 75%
```

```
value <- quantile(eliminated4$RI, probs = c(0.25,0.5,0.75))
print(value)
#breaks doesnt change much with outlier removal
eliminated4$cluster_scaledRI<-
cut(eliminated4$RI,breaks=c(0,1.5,2.1,2.9,800),labels=c("No risk","Slight risk" ,"Fair
risk","High Risk"),ordered=TRUE) #
str(eliminated4)
summary(eliminated4$cluster_scaledRI)
## dd is dataframe with cluster#Donot do k means
dd$cluster_scaledRI<-eliminated4$cluster_scaledRI
head(dd)
Also add RI and log RI so that it would become easier later.
dd$RI<-eliminated4$RI
dd$logRI<-eliminated4$logRI
head(dd)
summary(dd)
str(dd)
#removing outliers only after merging with dd to keep row numbers same
Q <- quantile(dd$RI, probs=c(.25, .75), na.rm = FALSE)
iqr <- IQR(dd$RI)
up <-  Q[2]+1.5*iqr # Upper Range
low<- Q[1]-1.5*iqr # Lower Range
dd<- subset(dd, dd$RI > (Q[1] - 1.5*iqr) & dd$RI < (Q[2]+1.5*iqr))
boxplot(dd)
summary(dd)
# calculate CDF to check whether out assumption was true or not
CDF <- ecdf(dd$RI)
# draw the cdf plot
plot( CDF )
#finding the values at 25%,40% and 75%
value <- quantile(dd$RI, probs = c(0.25,0.5,0.75))
print(value)
samplesize = 0.75*nrow(dd)
set.seed(1000)
```

```
index = sample(seq_len(nrow(dd)), size = samplesize)
#Creating training and test set
datatrain = dd[index,]
datatest = dd[-index,]
##########Ordinal logit modelling
https://www.youtube.com/watch?v=rrRrI9gElYA&t=404s
library(ordinal)
model1 <- clm(cluster~
Age+Wait_Time+Dry+Ped_Speed+LE_ILLE+Veh_typ+GR_Size+Dry,data =
datatrain,link = "logit")
summary(model1)
confint(model1)
exp(coef(model1))
exp(confint(model1))
# model afte rremovin insignifiacne varibales
model1 <- clm(cluster~
GR_Size+Wait_Time+Veh_gap+Veh_typ+Age+Gen+Ped_Speed+LE_ILLE,data =
datatrain,link = "logit")
summary(model1)
##with scaled RI
library(ordinal)
fin_model1 <-
clm(cluster_scaledRI~lane+Age+Veh_gap+LE_ILLE+Ped_Speed+Wait_Time,data =
datatrain)
summary(fin_model1)
model1 <-
clm(cluster_scaledRI~lane+Veh_gap+LE_ILLE+Ped_Speed+Veh_typ+Wait_Time,data
= datatrain)
summary(model1)
model_all <- clm(cluster_scaledRI
~lane+Veh_gap+Ped_Speed+Wait_Time+Gen+Veh_typ+GR_Size+LE_ILLE+Dry+Age,
data = datatrain)
summary(model_all)
#Stepwise regression in R both dierction stepwisemthod 2
```

```
#model is model with all variables
both_model <- step(model_all, direction = "both")
summary(both_model)
######use fin_model1 for sensitivity analysis#####
# Load necessary libraries
library(ordinal)  # For clm models
library(tidyr)    # For data reshaping
library(ggplot2)  # For plotting
# Step 1: Define the new data for prediction
predict_data <- data.frame(
  Wait_Time = seq(min(dd$Wait_Time[dd$Wait_Time>0], na.rm = TRUE),
                  max(dd$Wait_Time, na.rm = TRUE),
                  length.out = 100),  # Adjust range for your data
  lane = "1",                  # Fixed value, adjust as needed
  Age = "<20",                   # Fixed value, adjust as needed
  Veh_gap = mean(dd$Veh_gap, na.rm = TRUE),  # Mean of Veh_gap
  Ped_Speed = mean(dd$Ped_Speed, na.rm = TRUE),  # Mean of Ped_Speed
  LE_ILLE = "legal"              # Fixed value, adjust as needed
)
library(extrafont)
windowsFonts(Times = windowsFont("Times New Roman"))
# Ensure categorical variables match the model levels
predict_data$lane <- factor(predict_data$lane, levels = levels(dd$lane))
predict_data$Age <- factor(predict_data$Age, levels = levels(dd$Age))
predict_data$LE_ILLE <- factor(predict_data$LE_ILLE, levels = levels(dd$LE_ILLE))
# Step 2: Predict probabilities for each risk level
predicted_probs <- predict(fin_model1, newdata = predict_data, type = "prob")
# Step 3: Combine predicted probabilities with input data
# Convert the list of probabilities to a data frame
prob_matrix <- do.call(rbind, predicted_probs)
# Add Wait_Time to the data frame
predicted_probs_df <- cbind(predict_data, prob_matrix)
plot_data <- predicted_probs_df %>%
  pivot_longer(
```

```
    cols = c("No risk", "Slight risk", "Fair risk", "High Risk"),  # Specify exact column
names
    names_to = "Risk_Level",
    values_to = "Probability"
  )
# Rename Risk_Level values to remove "Pr()" if necessary
plot_data$Risk_Level <- gsub("Pr\\(", "", plot_data$Risk_Level)
plot_data$Risk_Level <- gsub("\\)", "", plot_data$Risk_Level)
# Step 5: Plot probabilities vs. Wait_Time
ggplot(plot_data, aes(x = Wait_Time, y = Probability, color = Risk_Level)) +
  geom_line(size = 1) +
  labs(
    title = "Risk Levels and Wait Time",
    x = "Wait Time (s)",
    y = "Probability",
    color = "Risk Level"
  ) +
   theme_minimal(base_size = 12) +
  theme(
    plot.title = element_text(family = "Times", size = 12, color = "black"),
    legend.title = element_text(family = "Times", size = 12, color = "black"),  # Try a
default font
    axis.title.x = element_text(family = "Times", size = 12, color = "black"),  # X-axis
label font
    axis.title.y = element_text(family = "Times", size = 12, color = "black"),
)
###############
#############Le_ILLe########
predict_data<-data.frame()
# Step 1: Define the new data for prediction
predict_data <- data.frame(
  LE_ILLE = c("legal","Illegal"),            # Fixed value, replace Wait_Time with
LE_ILLE
  lane = "1",                  # Fixed value, adjust as needed
```

```
  Age = "<20",                  # Fixed value, adjust as needed
  Veh_gap = mean(dd$Veh_gap, na.rm = TRUE),  # Mean of Veh_gap
  Ped_Speed = mean(dd$Ped_Speed, na.rm = TRUE),  # Mean of Ped_Speed
  Wait_Time = mean(dd$Wait_Time, na.rm = TRUE)

)
library(extrafont)
windowsFonts(Times = windowsFont("Times New Roman"))
# Ensure categorical variables match the model levels
predict_data$lane <- factor(predict_data$lane, levels = levels(dd$lane))
predict_data$Age <- factor(predict_data$Age, levels = levels(dd$Age))
predict_data$LE_ILLE <- factor(predict_data$LE_ILLE, levels = levels(dd$LE_ILLE))
# Step 2: Predict probabilities for each risk level
predicted_probs <- predict(fin_model1, newdata = predict_data, type = "prob")
# Step 3: Combine predicted probabilities with input data
# Convert the list of probabilities to a data frame
prob_matrix <- do.call(rbind, predicted_probs)
# Add LE_ILLE to the data frame
predicted_probs_df <- cbind(predict_data, prob_matrix)
plot_data <- predicted_probs_df %>%
  pivot_longer(
    cols = c("No risk", "Slight risk", "Fair risk", "High Risk"),  # Specify exact column
names
    names_to = "Risk_Level",
    values_to = "Probability"
  )
# Rename Risk_Level values to remove "Pr()" if necessary
plot_data$Risk_Level <- gsub("Pr\\(", "", plot_data$Risk_Level)
plot_data$Risk_Level <- gsub("\\)", "", plot_data$Risk_Level)
# Step 1: Check the structure of the 'predicted_probs_df'
str(predicted_probs_df)
# Step 2: Check the first few rows of 'plot_data'
head(plot_data)
# Step 3: Plot probabilities vs. LE_ILLE with points (no lines)
```

```
ggplot(plot_data, aes(x = LE_ILLE, y = Probability, color = Risk_Level)) +
  geom_point(size = 3) +  # Use points to visualize the data
  labs(
    title = "Risk Levels and LE_ILLE",
    x = "LE_ILLE",
    y = "Probability",
    color = "Risk Level"
  ) +
  theme_minimal(base_size = 12) +
  theme(
    plot.title = element_text(family = "Times", size = 12, color = "black"),
    legend.title = element_text(family = "Times", size = 12, color = "black"),  # Try a
default font
    axis.title.x = element_text(family = "Times", size = 12, color = "black"),  # X-axis
label font
    axis.title.y = element_text(family = "Times", size = 12, color = "black")
  )
###########simulated LE_ILLE############
# Load necessary libraries
library(ordinal)  # For clm models
library(tidyr)    # For data reshaping
library(ggplot2)  # For plotting
# Set a random seed for reproducibility
set.seed(123)
# Step 1: Simulate random values for the variables (using Monte Carlo-like sampling)
# Define the number of simulations (e.g., 1000 samples)
n_simulations <- 10000
# Calculate the rate parameter for the exponential distribution (lambda)
lambda_wait_time <- 1 / mean(dd$Wait_Time, na.rm = TRUE)  # Inverse of the mean of
Wait_Time
# Simulate random values for each variable
simulated_data <- data.frame(
  Wait_Time = rexp(n_simulations, rate = lambda_wait_time),  # Exponential distribution
for Wait_Time
```

```
  lane = sample(levels(dd$lane), n_simulations, replace = TRUE),  # Randomly sample from the levels of lane
  Age = sample(levels(dd$Age), n_simulations, replace = TRUE),  # Randomly sample from the levels of Age
  Veh_gap = rnorm(n_simulations, mean = mean(dd$Veh_gap, na.rm = TRUE), sd = sd(dd$Veh_gap, na.rm = TRUE)),  # Normal distribution for Veh_gap
  Ped_Speed = rnorm(n_simulations, mean = mean(dd$Ped_Speed, na.rm = TRUE), sd = sd(dd$Ped_Speed, na.rm = TRUE)),  # Normal distribution for Ped_Speed
  LE_ILLE = sample(levels(dd$LE_ILLE), n_simulations, replace = TRUE)  # Randomly sample from the levels of LE_ILLE
)
# Step 2: Use the simulated data for prediction
# Ensure categorical variables match the model levels
simulated_data$lane <- factor(simulated_data$lane, levels = levels(dd$lane))
simulated_data$Age <- factor(simulated_data$Age, levels = levels(dd$Age))
simulated_data$LE_ILLE <- factor(simulated_data$LE_ILLE, levels = levels(dd$LE_ILLE))
# Predict probabilities for each risk level based on the simulated data
predicted_probs <- predict(fin_model1, newdata = simulated_data, type = "prob")
# Step 3: Combine predicted probabilities with the simulated input data
# Convert the list of probabilities to a data frame
prob_matrix <- do.call(rbind, predicted_probs)
# Add the simulated variables to the data frame
simulated_probs_df <- cbind(simulated_data, prob_matrix)
# Step 4: Reshape the data for plotting
plot_data <- simulated_probs_df %>%
  pivot_longer(
    cols = c("No risk", "Slight risk", "Fair risk", "High Risk"),  # Specify exact column names
    names_to = "Risk_Level",
    values_to = "Probability"
  )
plot_data$Risk_Level <- factor(plot_data$Risk_Level,
                               levels = c("No risk", "Slight risk", "Fair risk", "High Risk"))
```

```
### Box Plot for Probability Distribution across LE_ILLE
ggplot(plot_data, aes(x = LE_ILLE, y = Probability, fill = Risk_Level)) +
  geom_boxplot(outlier.shape = NA, alpha = 0.5) +  # Boxplot to show the distribution of probabilities
  labs(
    title = "Risk Levels and LE_ILLE (Box Plot)",
    x = "LE_ILLE",
    y = "Probability",
    fill = "Risk Level"
  ) +
  theme_minimal(base_size = 12) +
  theme(
    plot.title = element_text(family = "Times", size = 12, color = "black"),
    axis.title.x = element_text(family = "Times", size = 12, color = "black"),
    axis.title.y = element_text(family = "Times", size = 12, color = "black"),
    legend.title = element_text(family = "Times", size = 12, color = "black")
  )
### Distribution Plot (Density) for Each Risk Level
library(ggplot2)
# Assuming plot_data contains the predicted probabilities for both 'legal' and 'illegal'
ggplot(plot_data, aes(x = Probability, fill = Risk_Level, color = Risk_Level)) +
  geom_density(alpha = 0.4) +  # Density plot with transparency
  facet_wrap(~ LE_ILLE, scales = "free") +  # Separate the plots by LE_ILLE (Legal/Illegal)
  labs(
    title = "Density Plot of Predicted Probabilities by LE_ILLE and Risk Level",
    x = "Predicted Probability",
    y = "Density"
  ) +
  theme_minimal(base_size = 12) +
  theme(
    plot.title = element_text(family = "Times", size = 12, color = "black"),
    axis.title.x = element_text(family = "Times", size = 12, color = "black"),
    axis.title.y = element_text(family = "Times", size = 12, color = "black"),
```

```
    legend.title = element_text(family = "Times", size = 12, color = "black")
  )
############################
#####################wait_time
# Step 1: Define the new data for prediction
wait_time_seq <- seq(
  min(dd$Wait_Time[dd$Wait_Time >= 0], na.rm = TRUE),  # Filter out negative values
  max(dd$Wait_Time, na.rm = TRUE),
  length.out = 100
)
library(extrafont)
windowsFonts(Times = windowsFont("Times New Roman"))
predict_data <- expand.grid(
  LE_ILLE = c("legal", "Illegal"),            # Fixed value
  lane = "1",                              # Fixed value
  Age = "<20",                               # Fixed value
  Veh_gap = mean(dd$Veh_gap, na.rm = TRUE),    # Mean of Veh_gap
  Ped_Speed = mean(dd$Ped_Speed, na.rm = TRUE),# Mean of Ped_Speed
  Wait_Time = wait_time_seq                   # Sequence for Wait_Time
)
# Ensure categorical variables match the model levels
predict_data$lane <- factor(predict_data$lane, levels = levels(dd$lane))
predict_data$Age <- factor(predict_data$Age, levels = levels(dd$Age))
predict_data$LE_ILLE <- factor(predict_data$LE_ILLE, levels = levels(dd$LE_ILLE))
# Step 2: Predict probabilities for each risk level
predicted_probs <- predict(fin_model1, newdata = predict_data, type = "prob")
# Step 3: Combine predicted probabilities with input data
prob_matrix <- do.call(rbind, predicted_probs) # Convert list of probabilities to a matrix
predicted_probs_df <- cbind(predict_data, prob_matrix)
# Step 4: Reshape data for plotting
library(tidyr)
plot_data <- predicted_probs_df %>%
  pivot_longer(
```

```
    cols = c("No risk", "Slight risk", "Fair risk", "High Risk"),  # Specify exact column
names
    names_to = "Risk_Level",
    values_to = "Probability"
  )
# Reorder Risk_Level for legend order
plot_data$Risk_Level <- factor(plot_data$Risk_Level, levels = c("No risk", "Slight risk",
"Fair risk", "High Risk"))
# Step 5: Plot probabilities vs. Wait_Time
library(ggplot2)
ggplot(plot_data, aes(x = Wait_Time, y = Probability, color = Risk_Level, linetype =
LE_ILLE)) +
  geom_line(size = 1) +
  labs(
    title = "Predicted Probabilities Across Wait Time",
    x = "Wait Time",
    y = "Probability",
    color = "Risk Level",
    linetype = "LE_ILLE"
  ) +
  theme_minimal(base_size = 12) +
  theme(
    plot.title = element_text(family = "Times", size = 12, color = "black"),
    legend.title = element_text(family = "Times", size = 12, color = "black"),  # Try a
default font
    axis.title.x = element_text(family = "Times", size = 12, color = "black"),  # X-axis
label font
    axis.title.y = element_text(family = "Times", size = 12, color = "black"),
)
##############
```